P&S Modern SSDs

NAND Flash Read/Write Operations

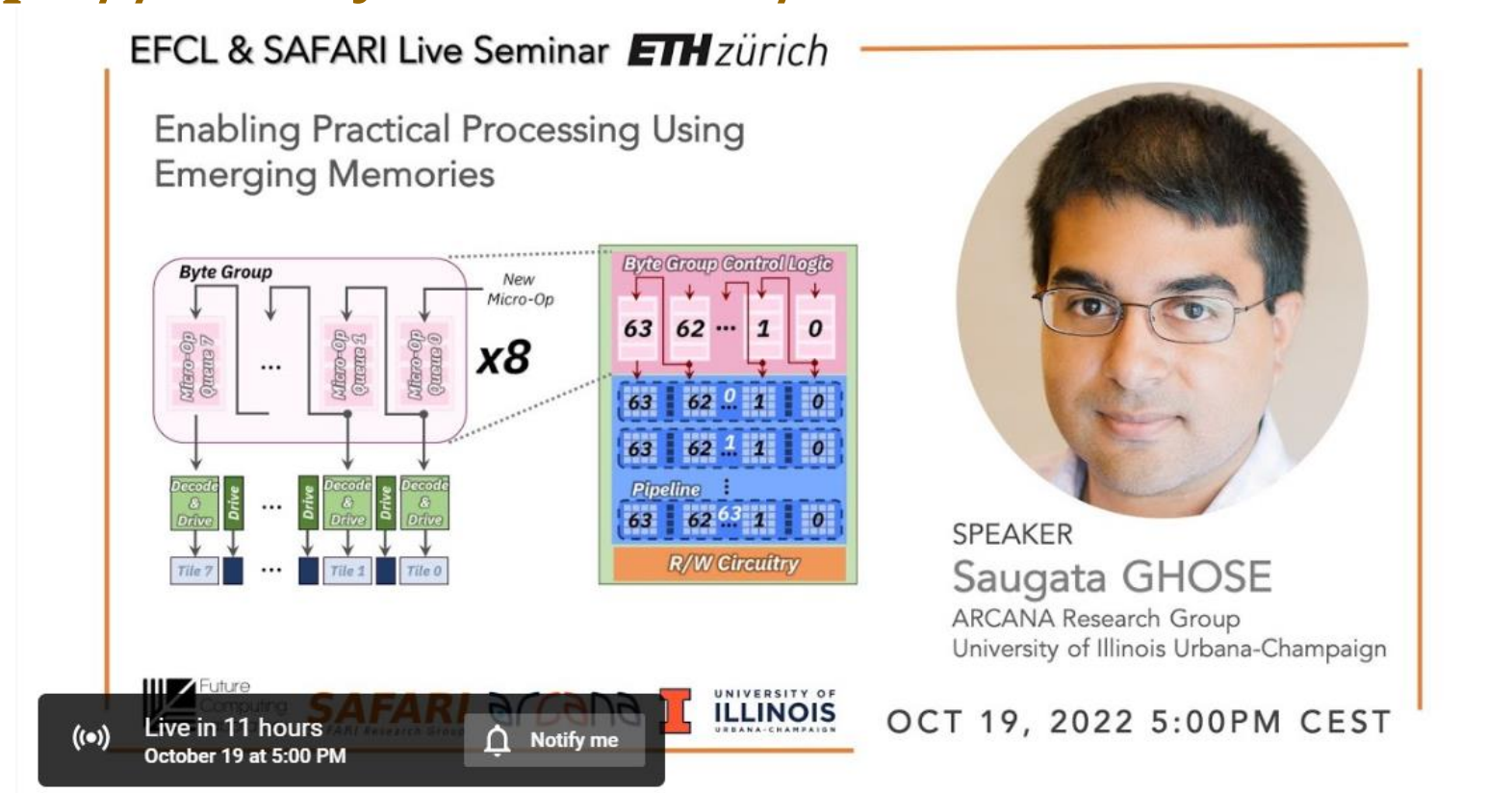
Dr. Mohammad Sadrosadati
Prof. Onur Mutlu
ETH Zürich

Fall 2022

19 October 2022

Upcoming EFCL & SAFARI Live Seminar (Oct. 19)

https://www.youtube.com/watch?v=5BRlkCCL-hY



Today's Agenda

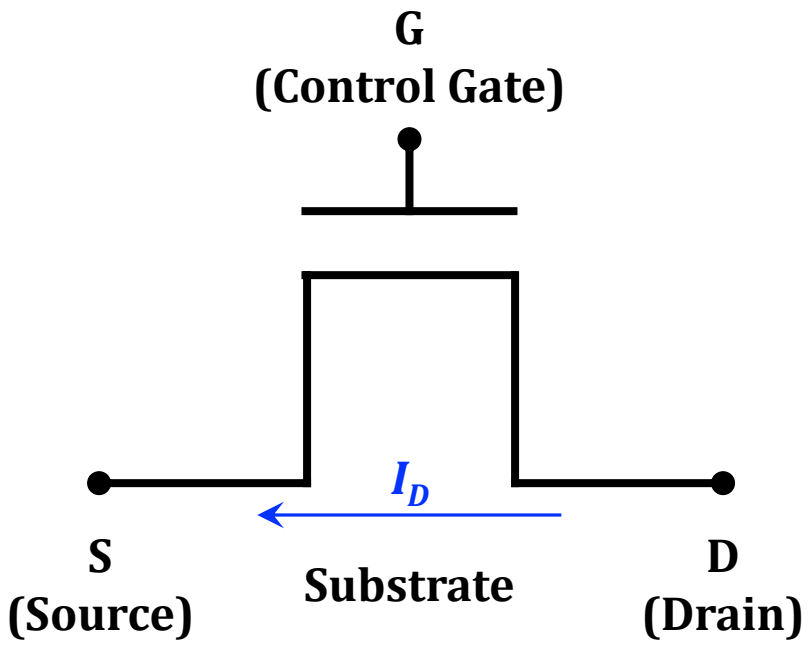
- NAND Flash Page Read/Write
- Sensing Circuitry

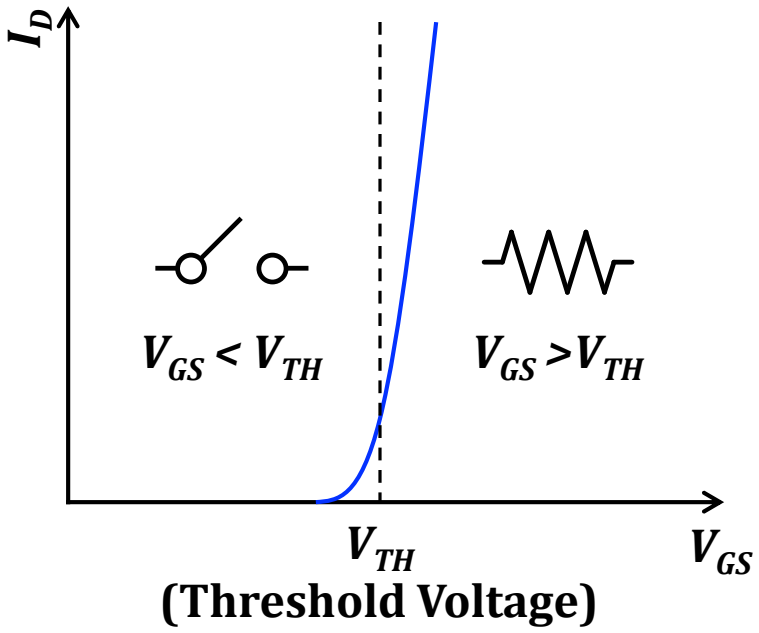
Today's Agenda

- NAND Flash Page Read/Write
- Sensing Circuitry

A Flash Cell

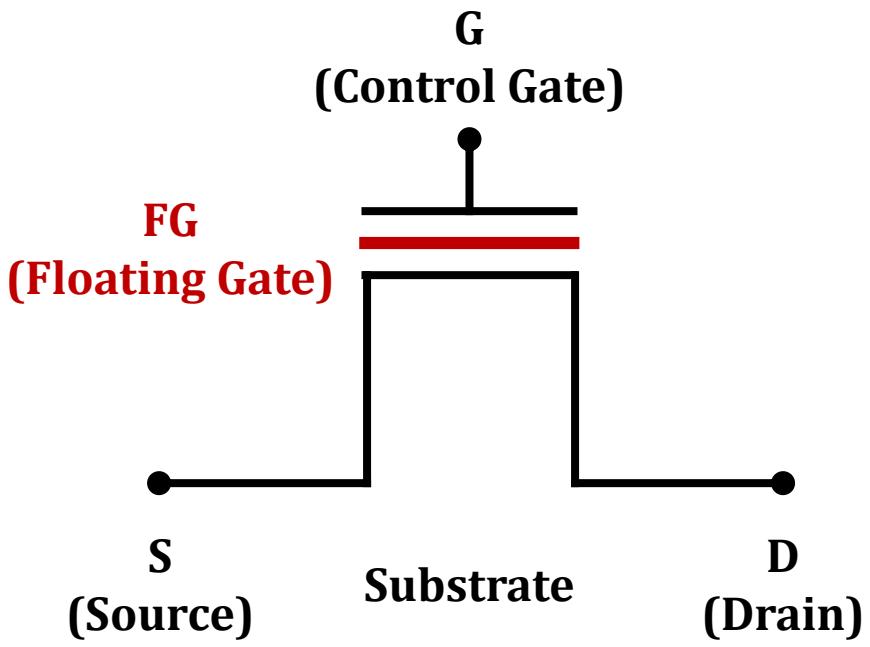
- Basically, it is a transistor

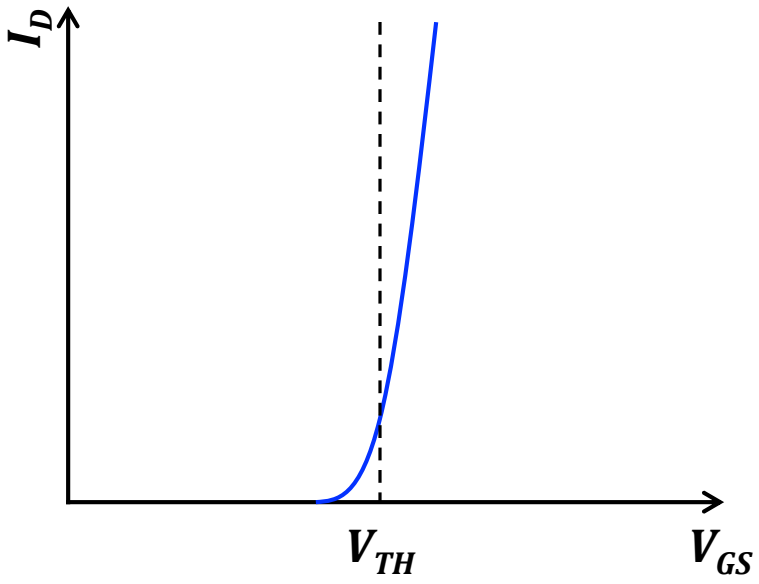




A Flash Cell

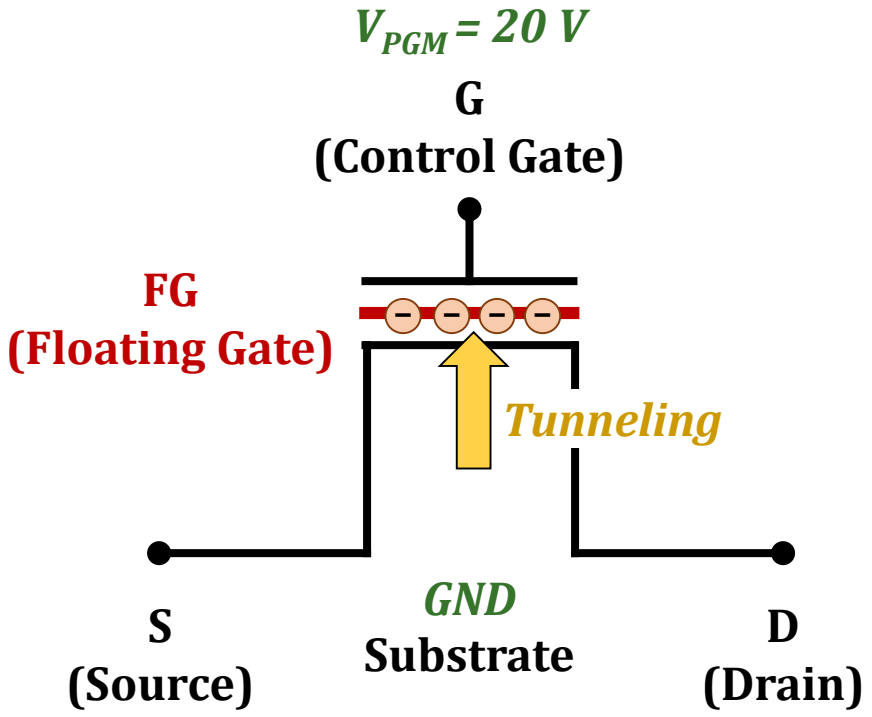
- Basically, it is a transistor
  - w/ a special material: Floating gate (2D) or Charge trap (3D)

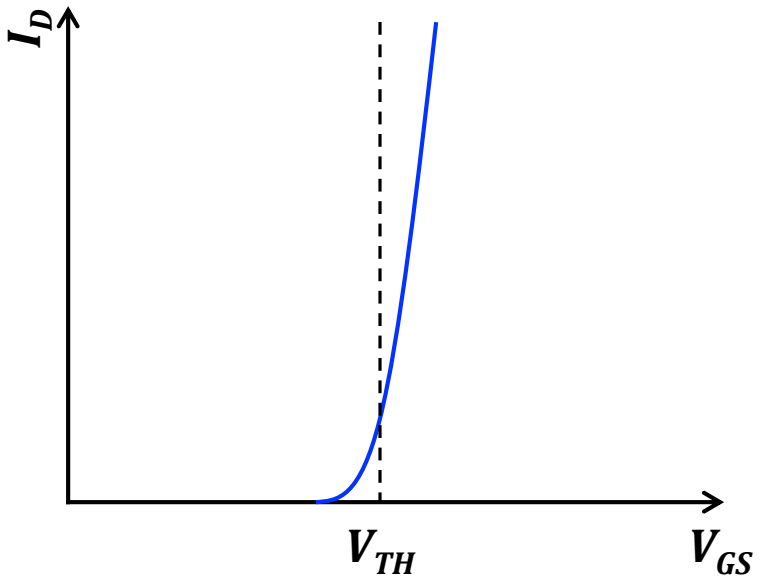




A Flash Cell

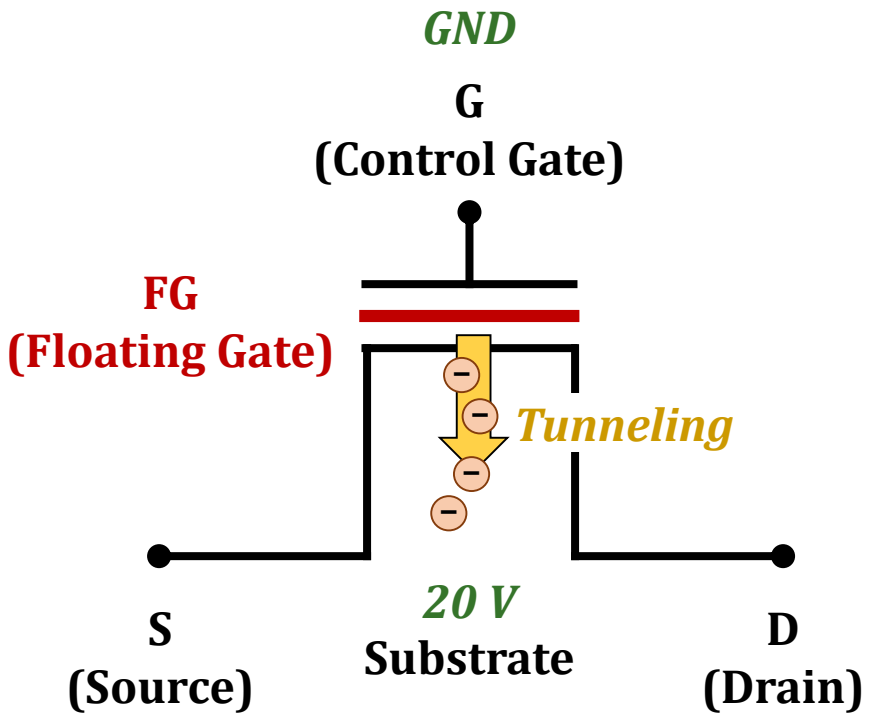
- Basically, it is a transistor
  - w/ a special material: Floating gate (2D) or Charge trap (3D)
  - Can hold electrons in a non-volatile manner

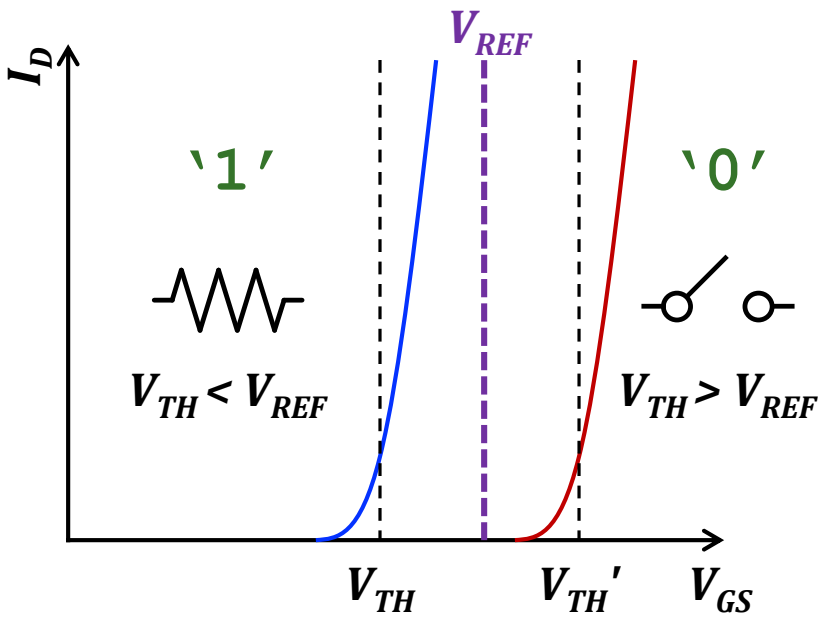




A Flash Cell

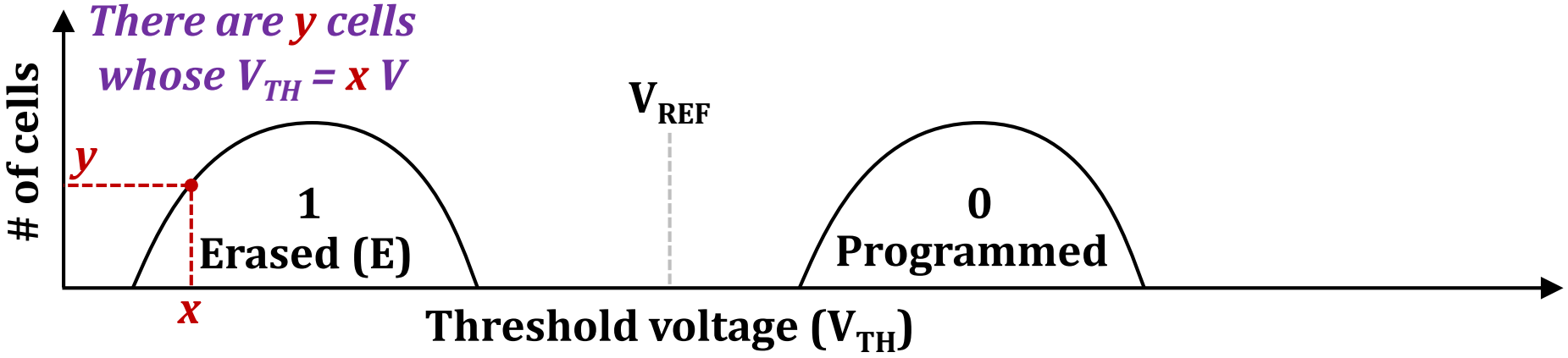
- Basically, it is a transistor
  - w/ a special material: Floating gate (2D) or Charge trap (3D)
  - Can hold electrons in a non-volatile manner
  - Changes the cell's threshold voltage (V_{TH})





Threshold Voltage Distribution

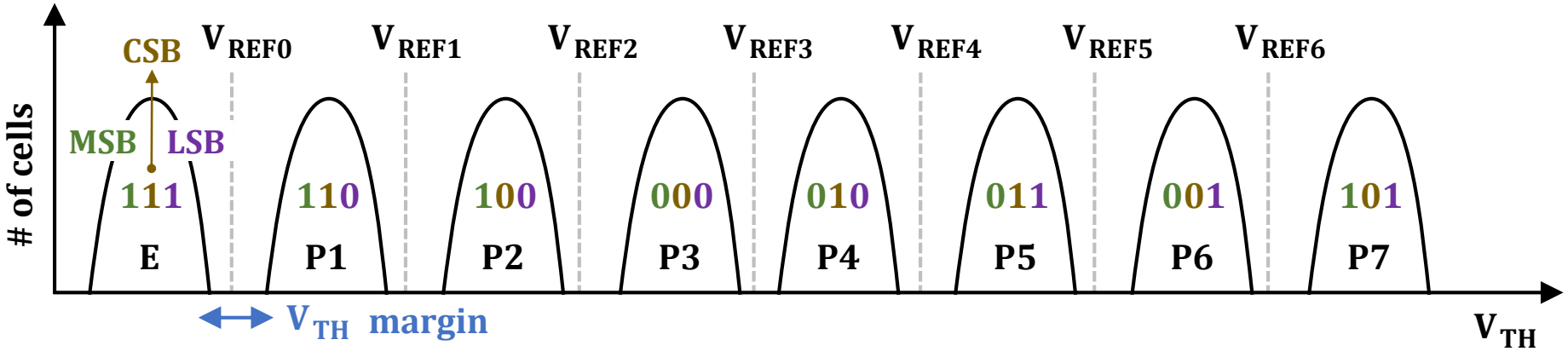
- V_{TH} distribution of cells in a programmed page/block/chip



- Why distribution? Variations across the cells
  - Some cells are more easily programmed or erased

V_{TH} Distribution of MLC NAND Flash

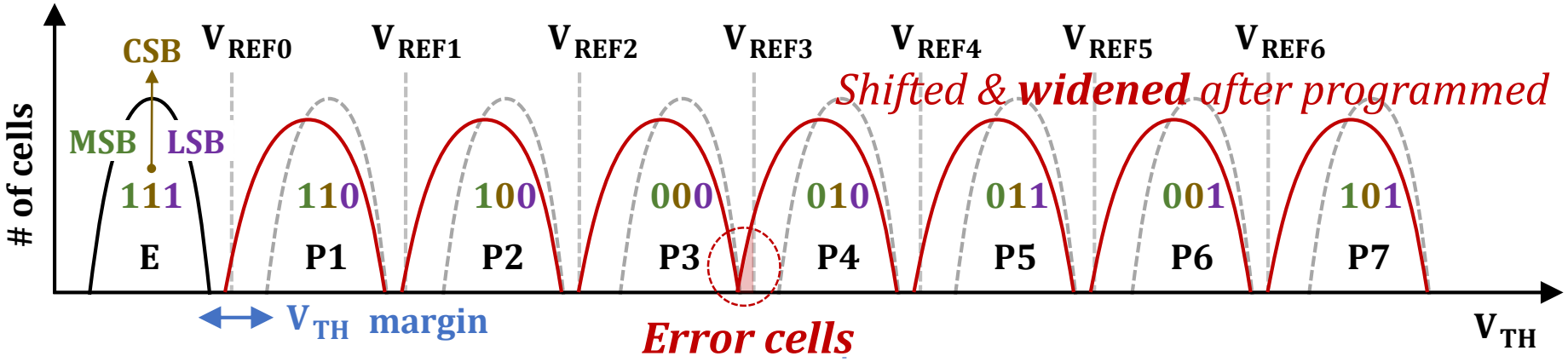
- Multi-level cell (MLC) technique
  - 2^m V_{TH} states required to store *m* bits in a single flash cell



- Limited width of the V_{TH} window: Need to
  - Make each V_{TH} state narrow
  - Guarantee sufficient margins b/w adjacent V_{TH} states

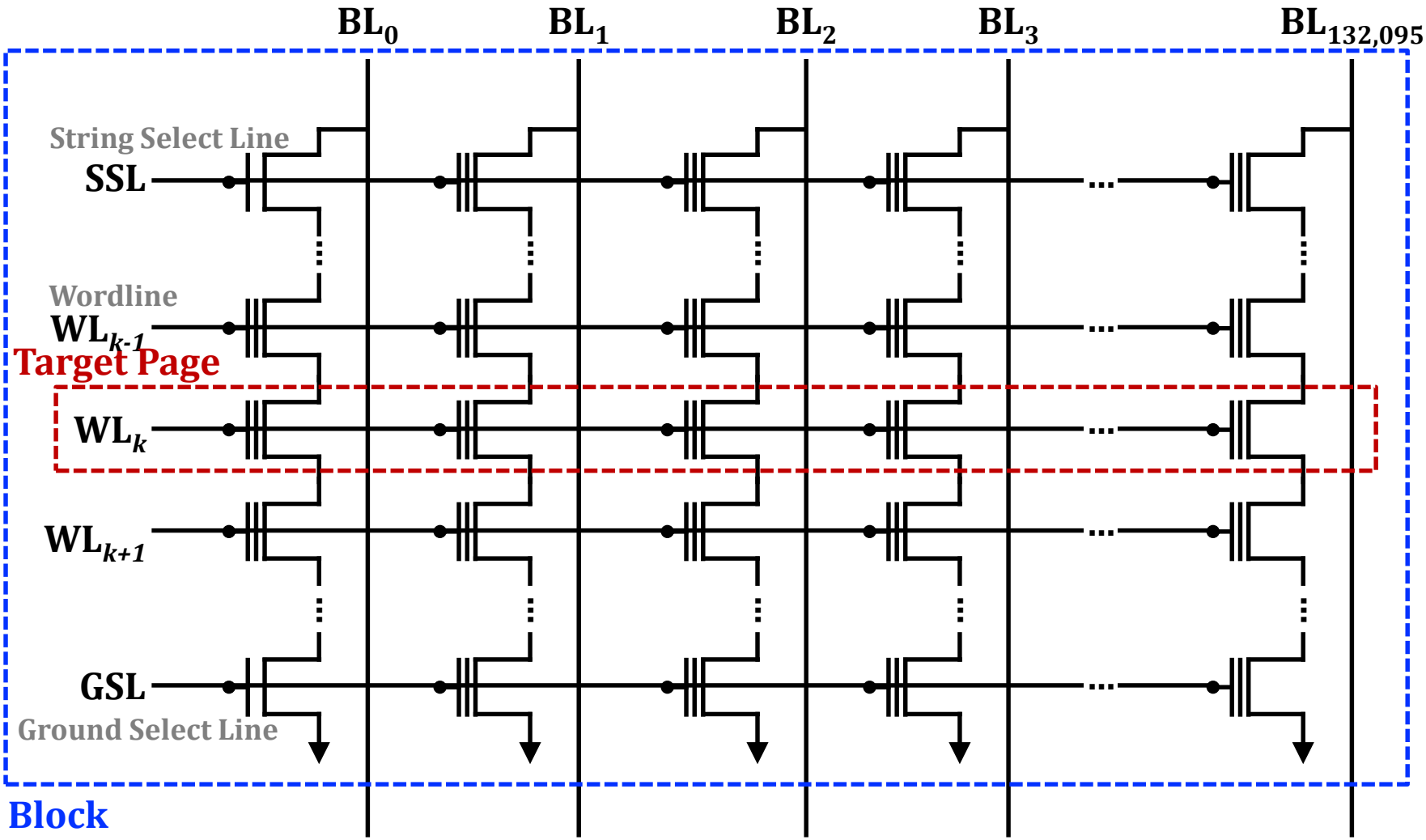
V_{TH} Distribution of MLC NAND Flash

- Multi-level cell (MLC) technique
  - 2^m V_{TH} states required to store *m* bits in a single flash cell



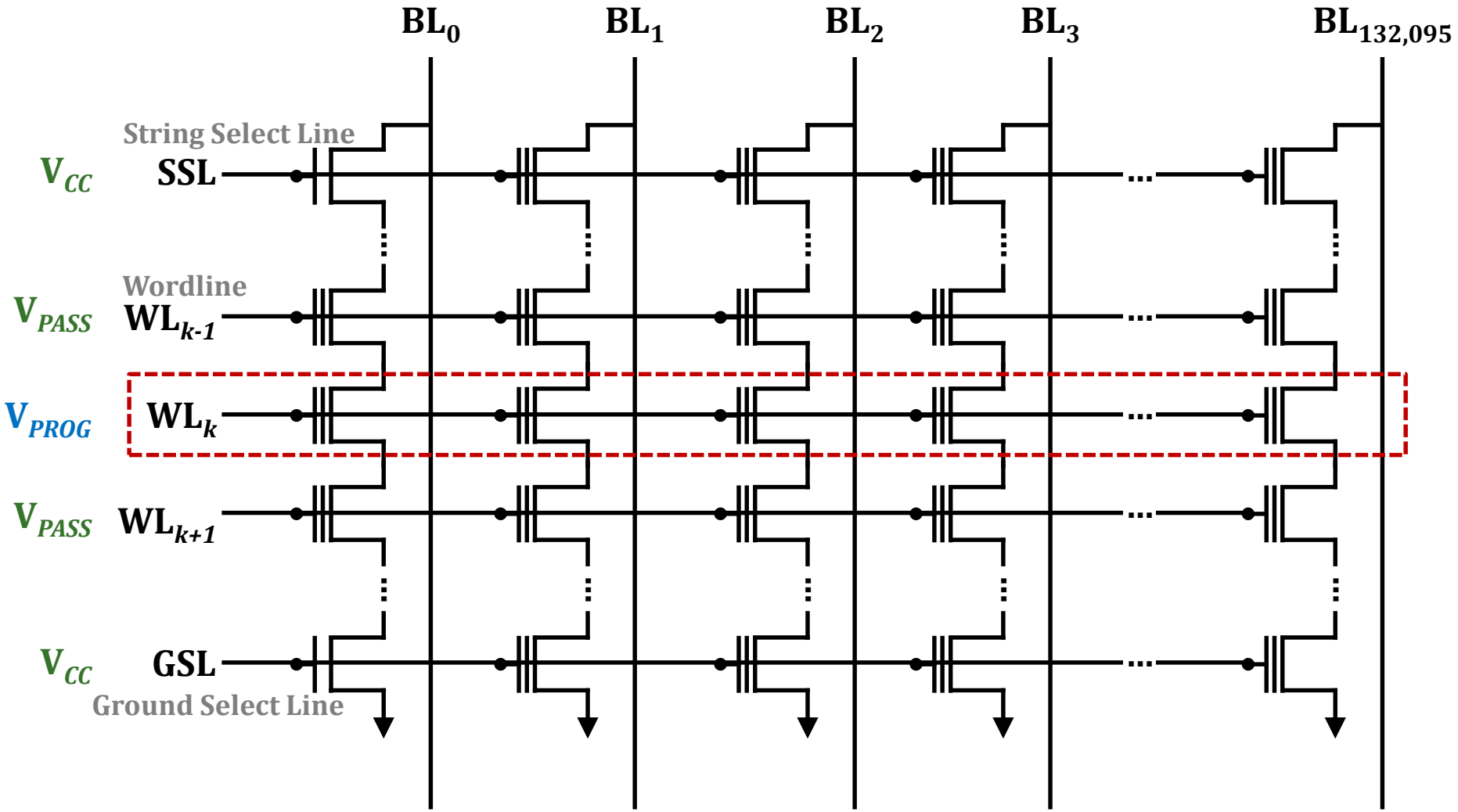
- Limited width of the V_{TH} window: Need to
  - Make each V_{TH} state narrow
  - Guarantee sufficient margins b/w adjacent V_{TH} states
    - V_{TH} changes over time after programmed
    - Narrower margins → Lower reliability
    - More bits per cell → higher density but lower reliability

Basic Operation: Page Program



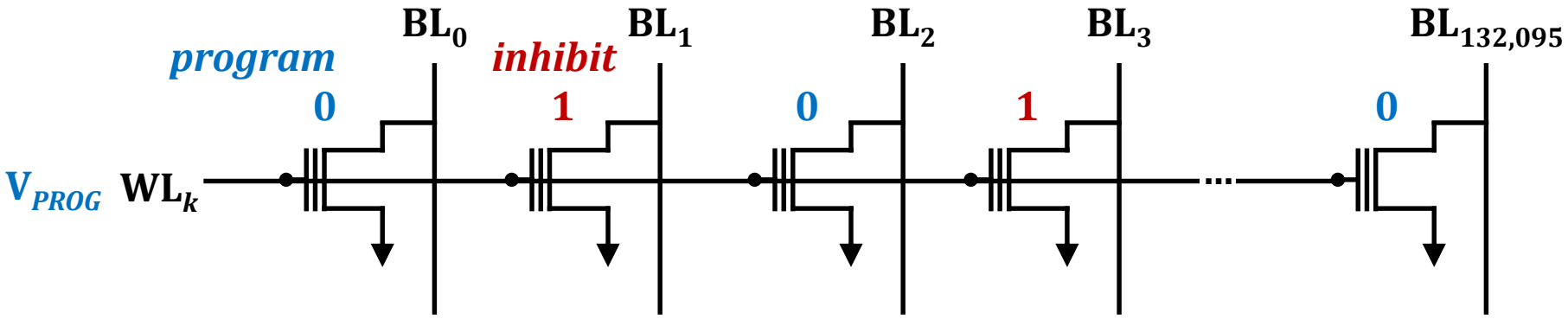
Basic Operation: Page Program

- WL control – All other cells operate as a resistance



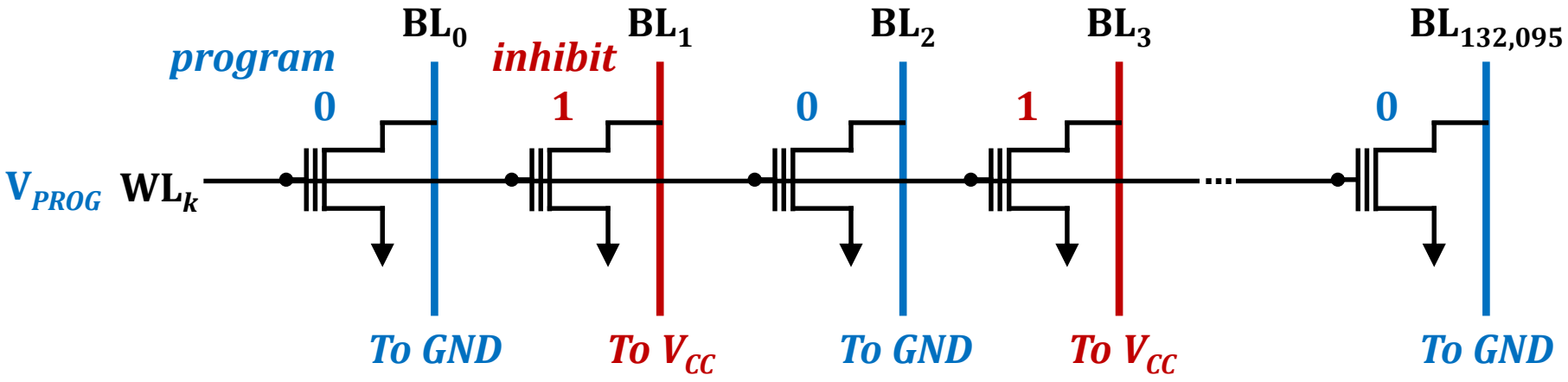
Basic Operation: Page Program

- BL control – Inhibits cells to not be programmed

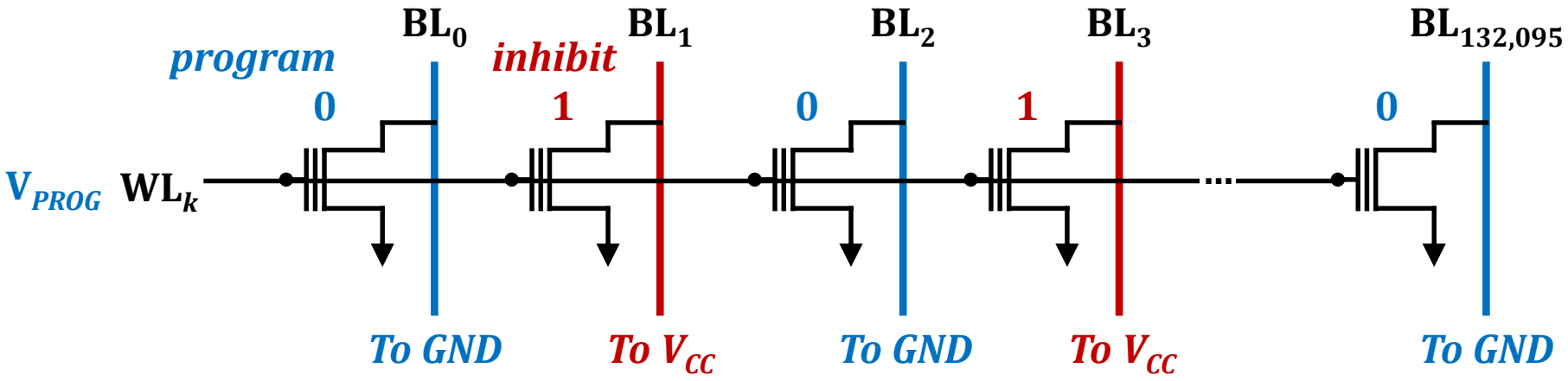


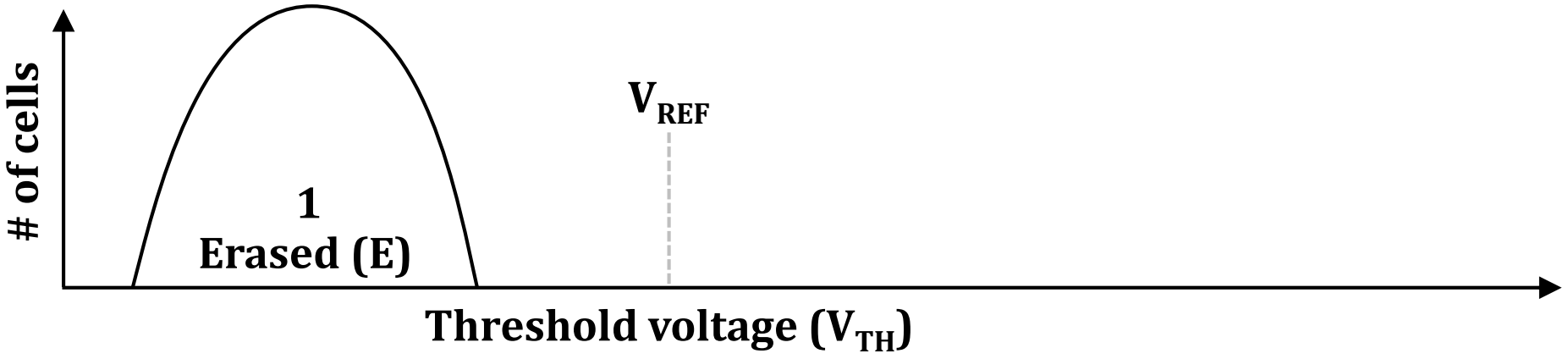
Basic Operation: Page Program

- BL control – Inhibits cells to not be programmed

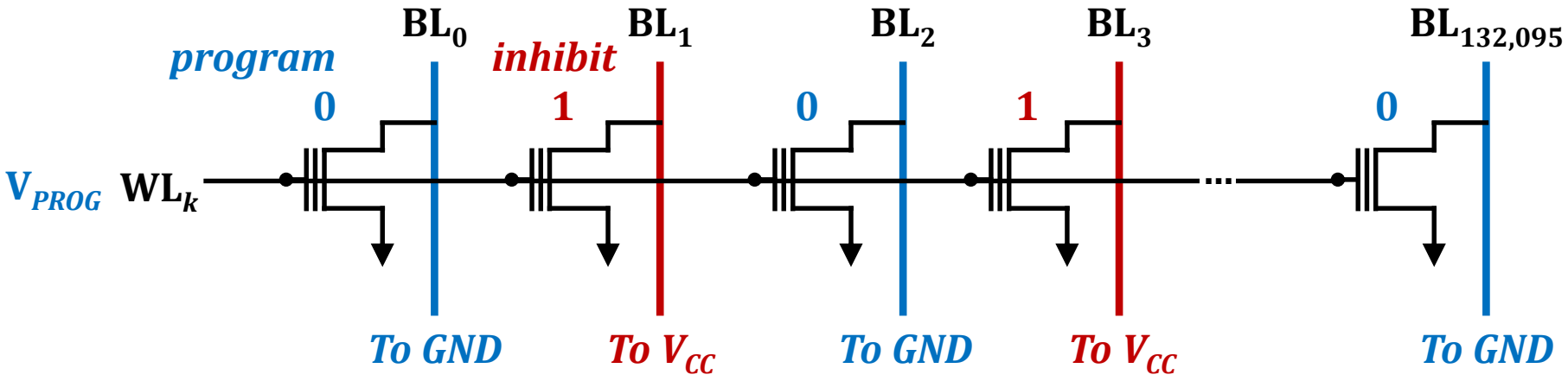


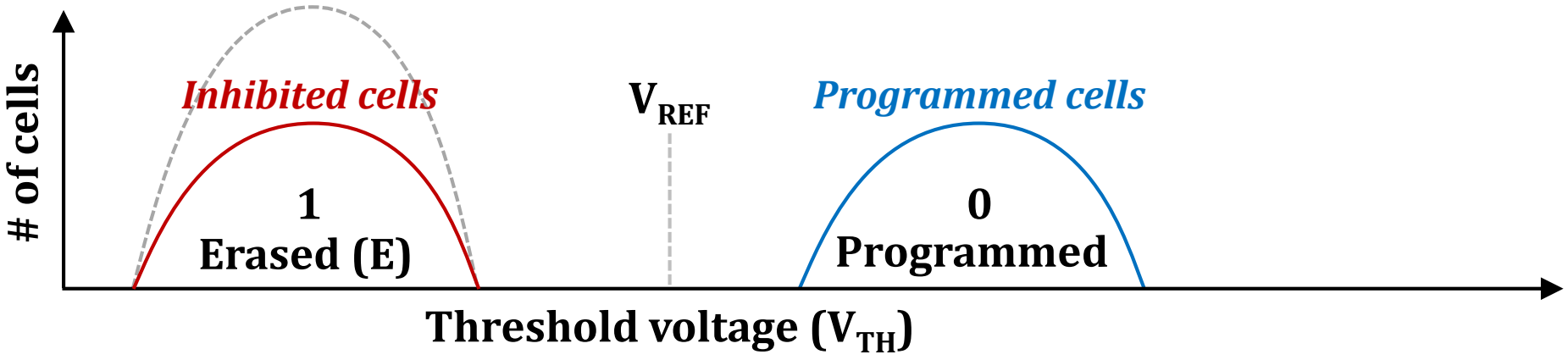
Basic Operation: Page Program



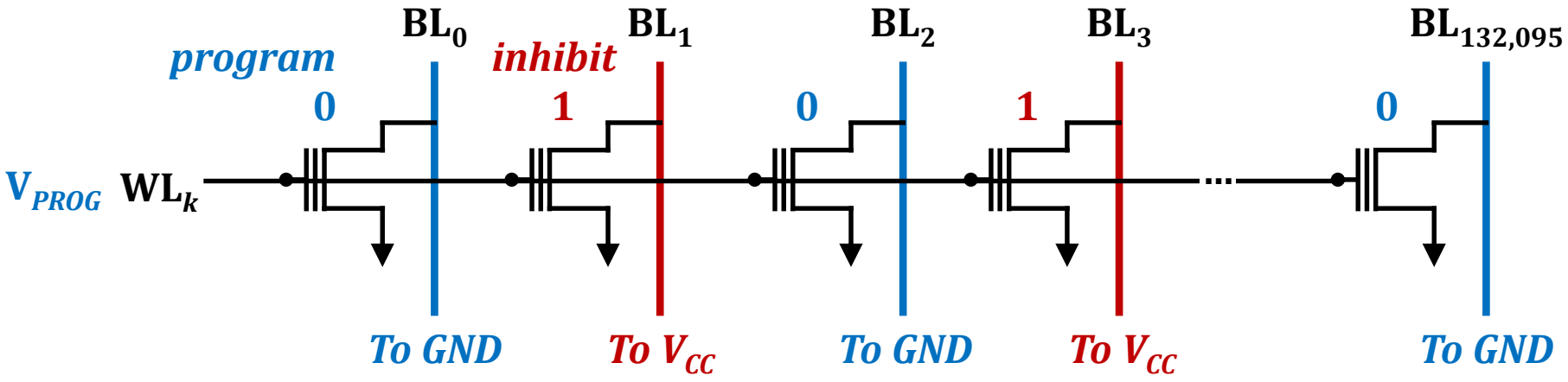


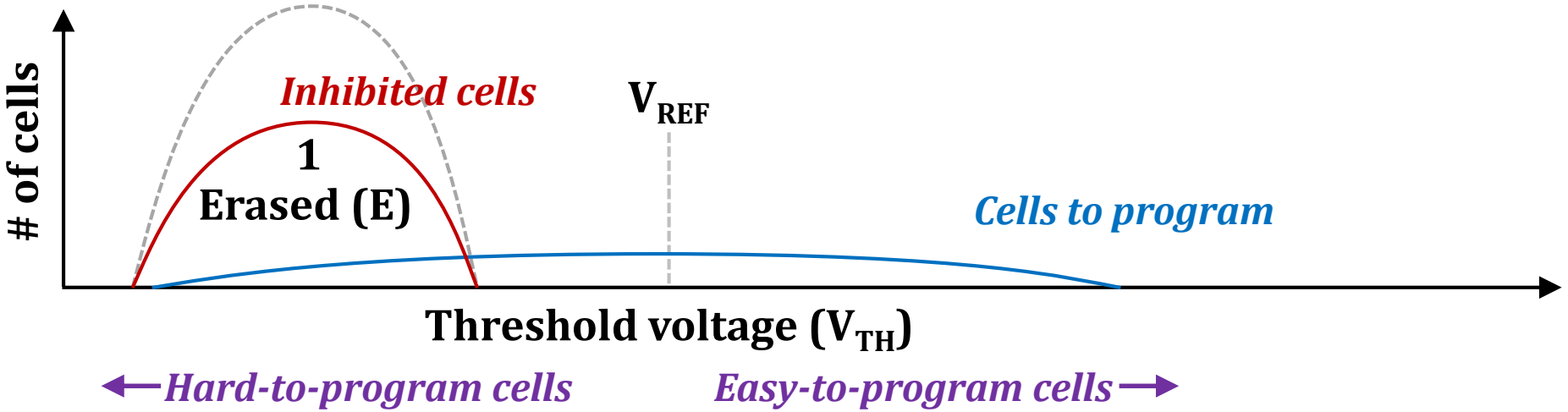
Basic Operation: Page Program





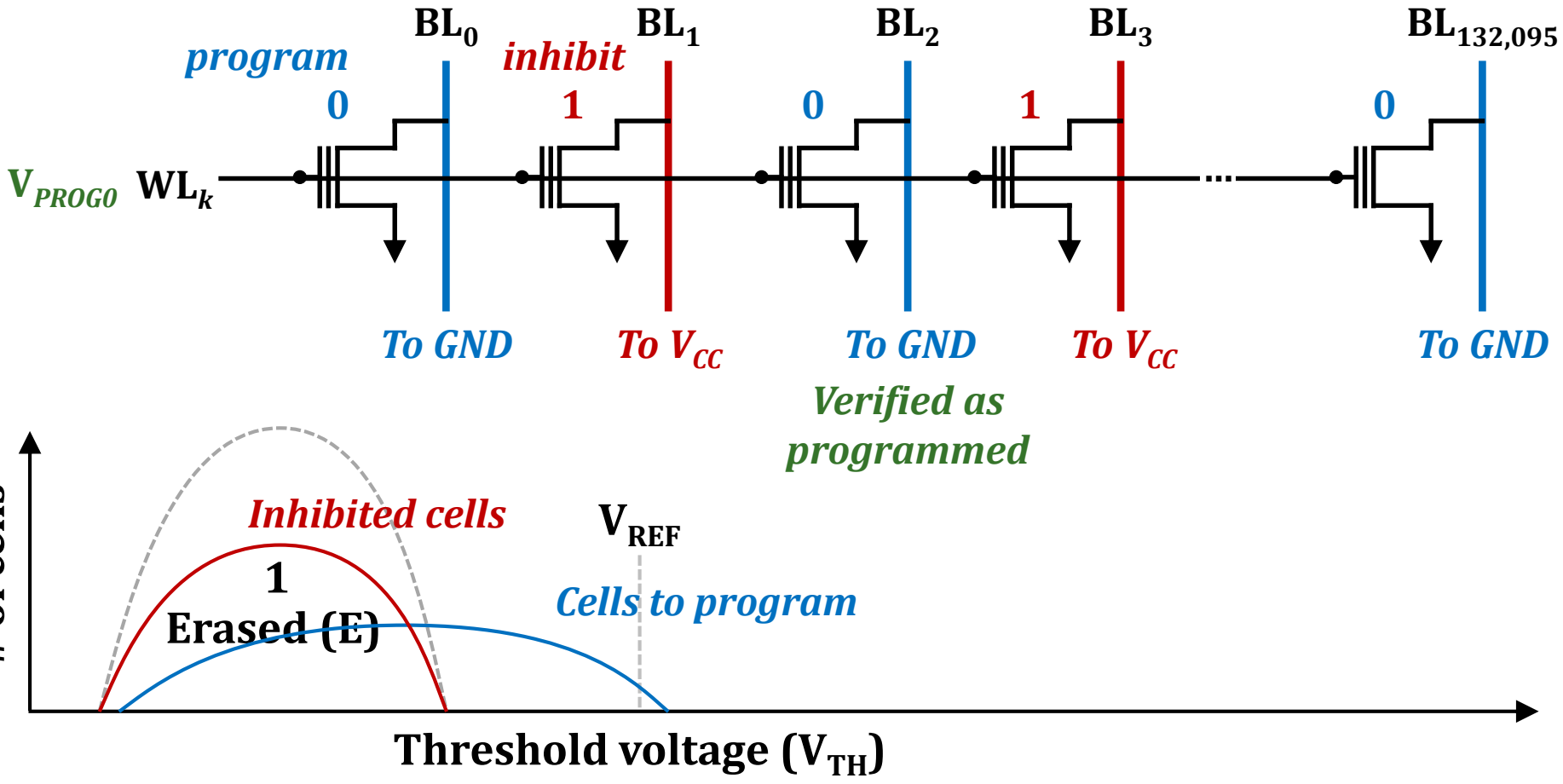
Basic Operation: Page Program





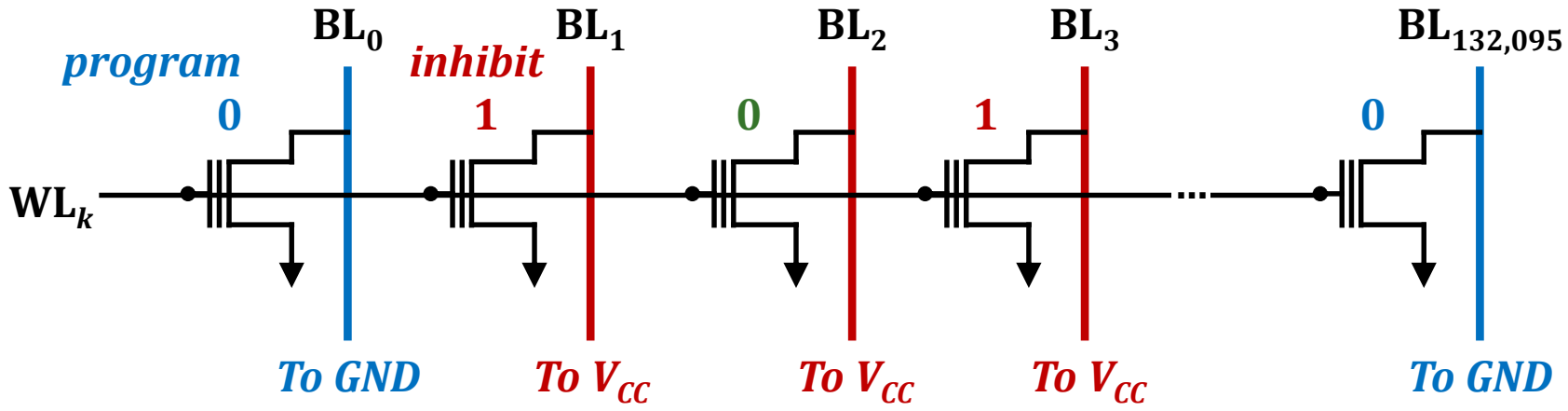
Basic Operation: Page Program

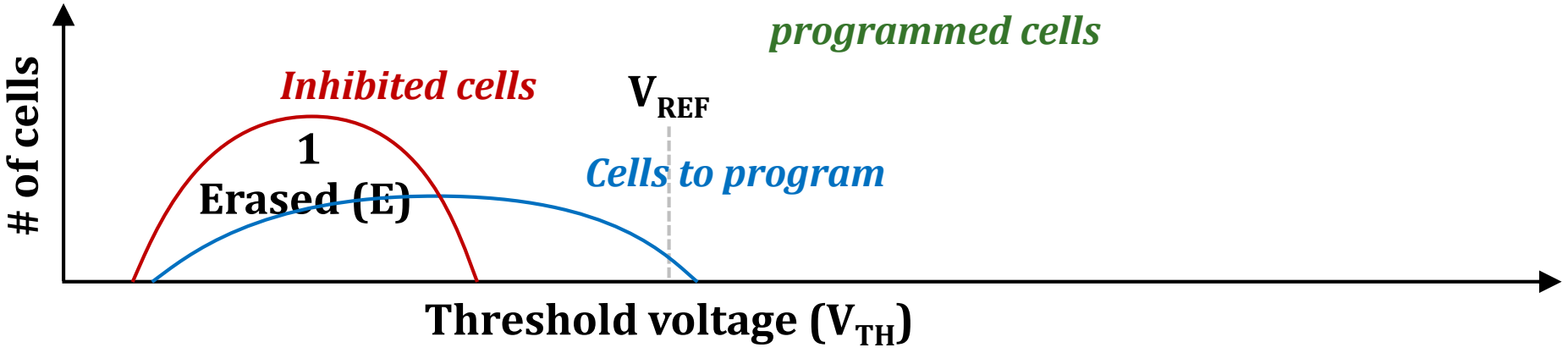
- Incremental Step-Pulse Programming (ISPP)



Basic Operation: Page Program

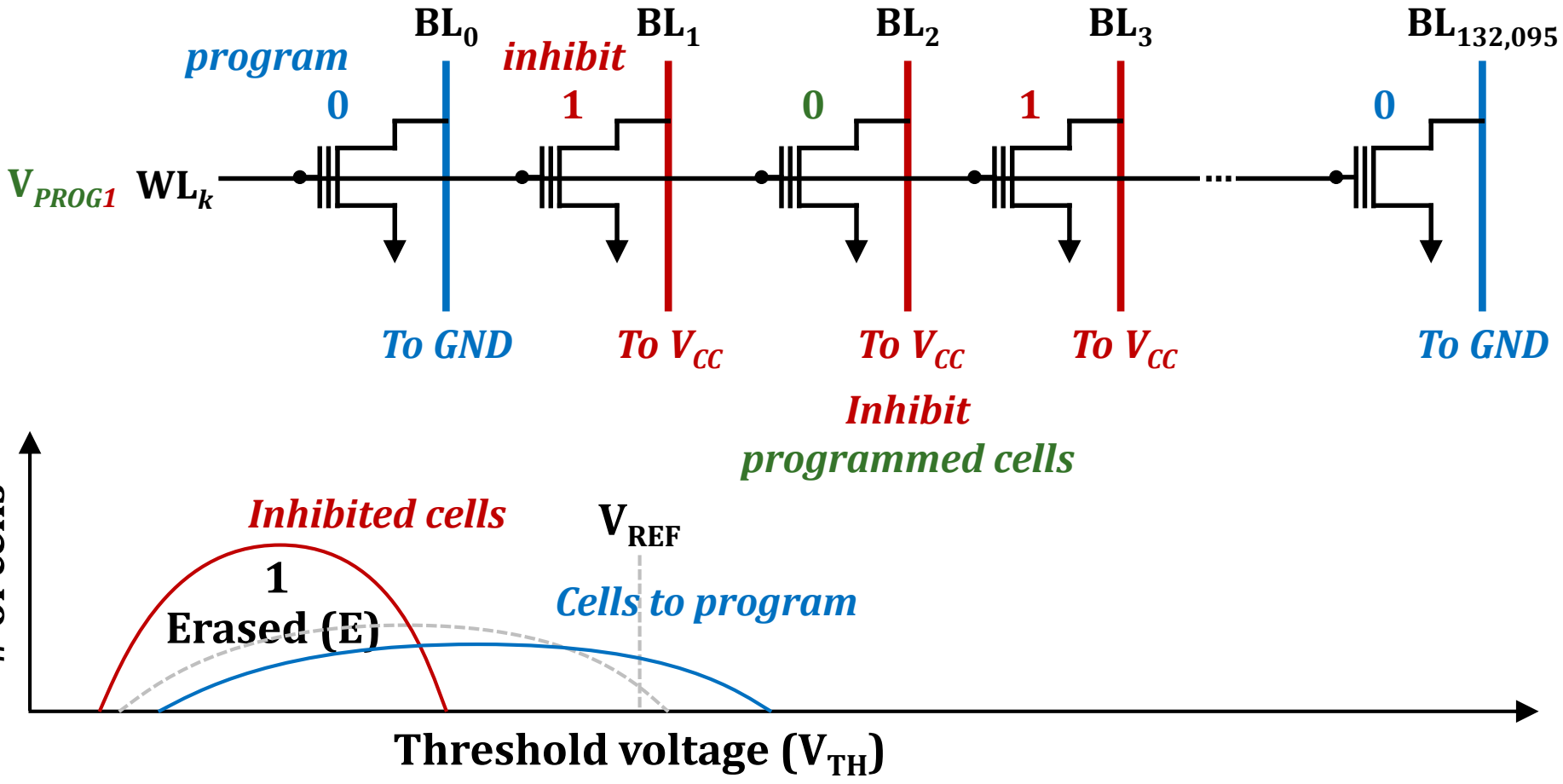
- Incremental Step-Pulse Programming (ISPP)





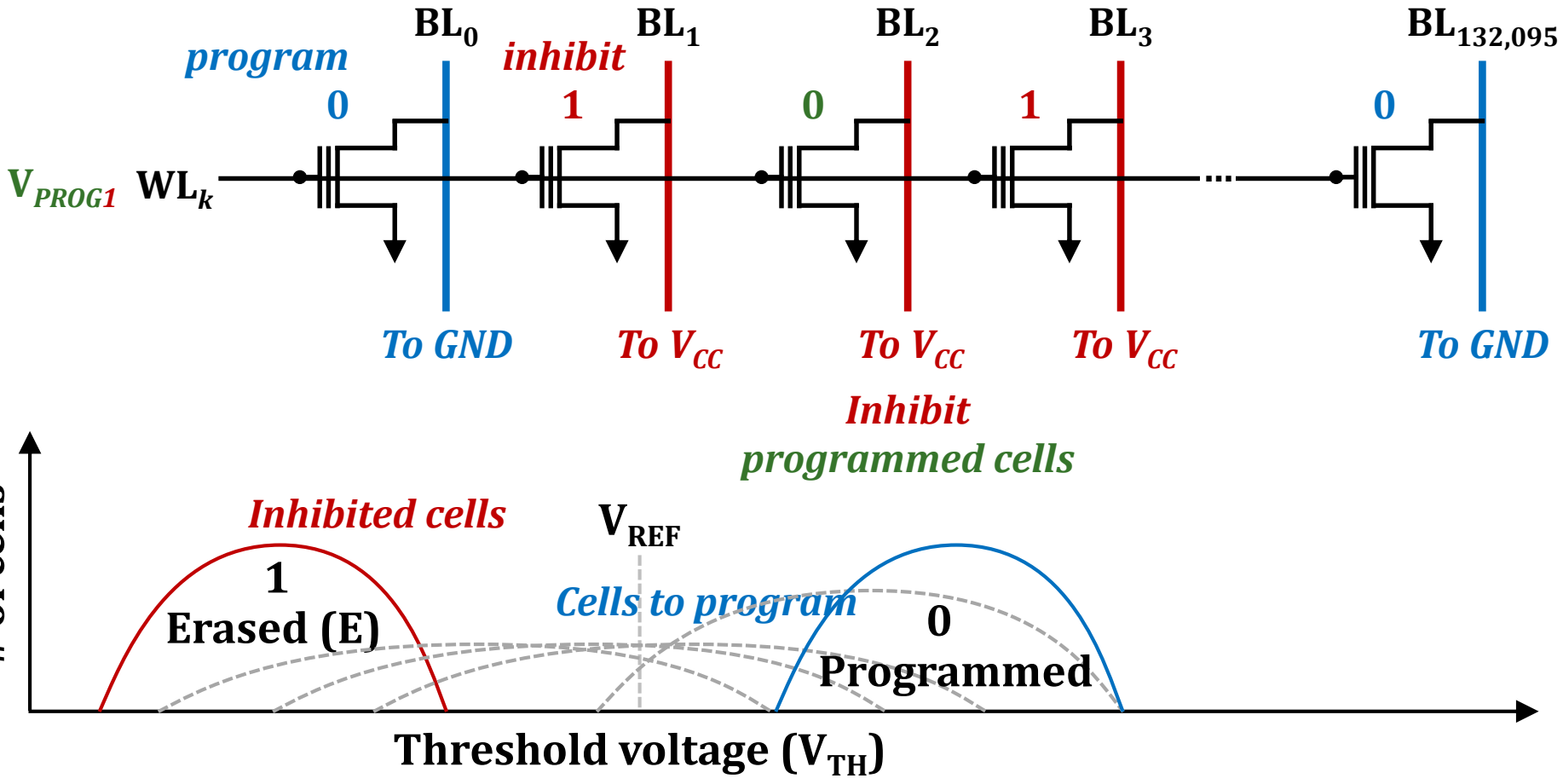
Basic Operation: Page Program

- Incremental Step-Pulse Programming (ISPP)



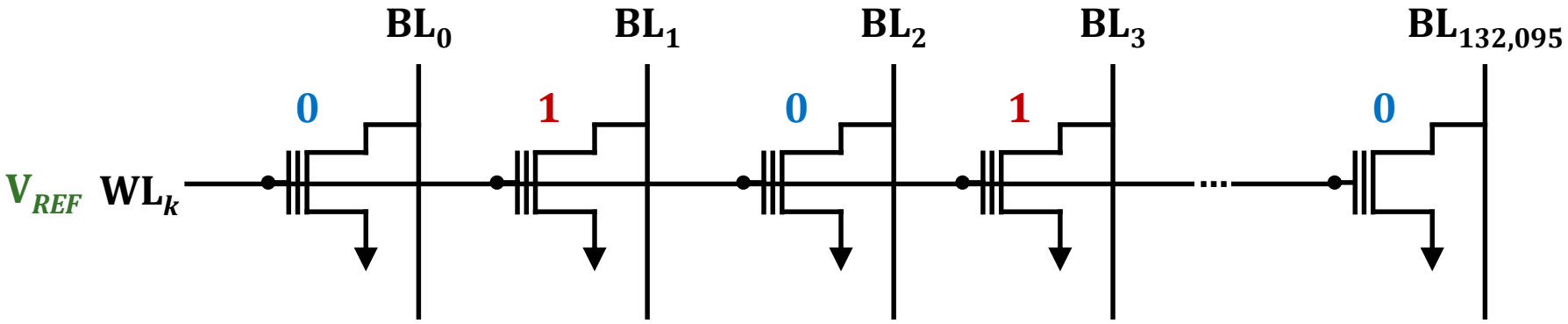
Basic Operation: Page Program

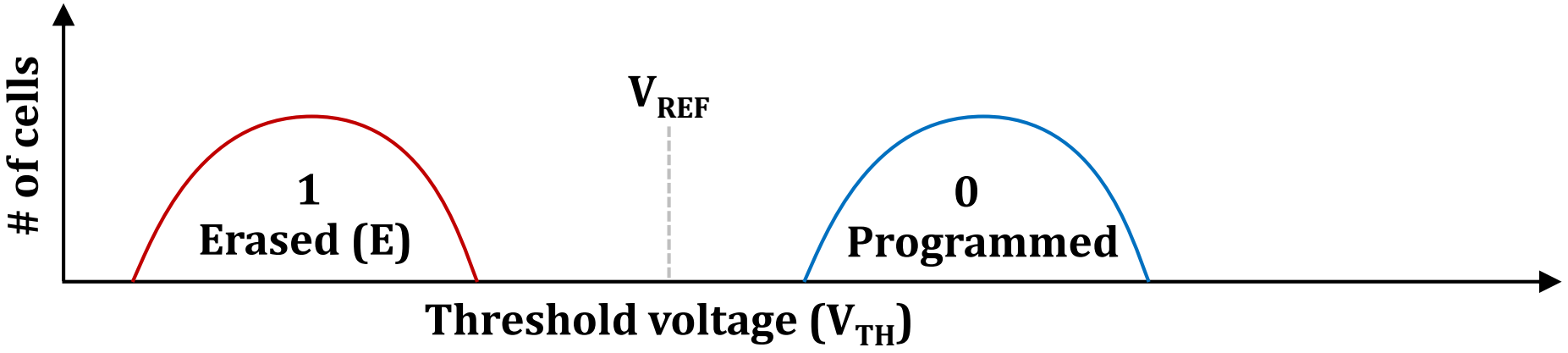
- Incremental Step-Pulse Programming (ISPP)



Basic Operation: Page Read

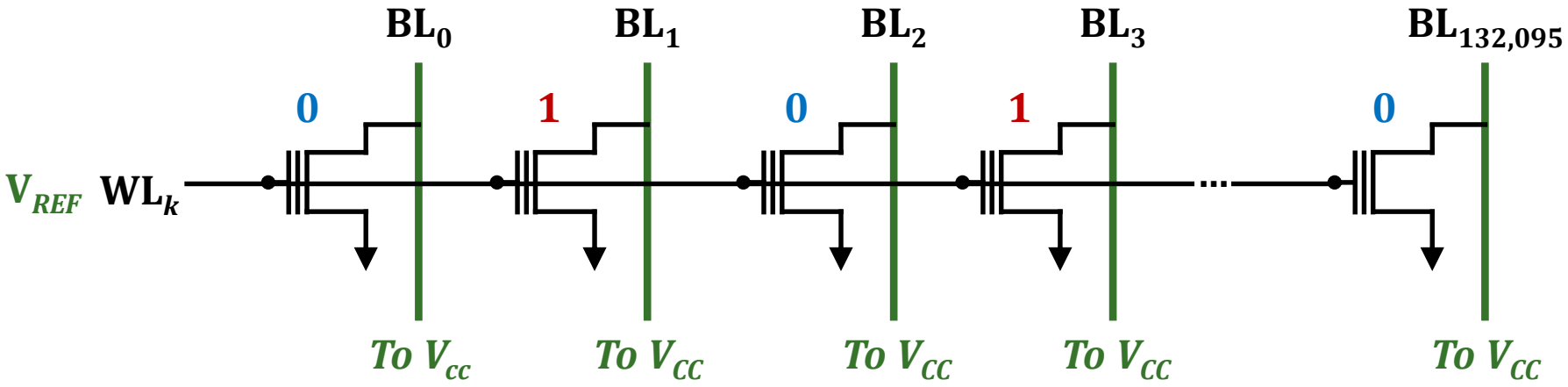
- WL control – All other cells operate as a resistance

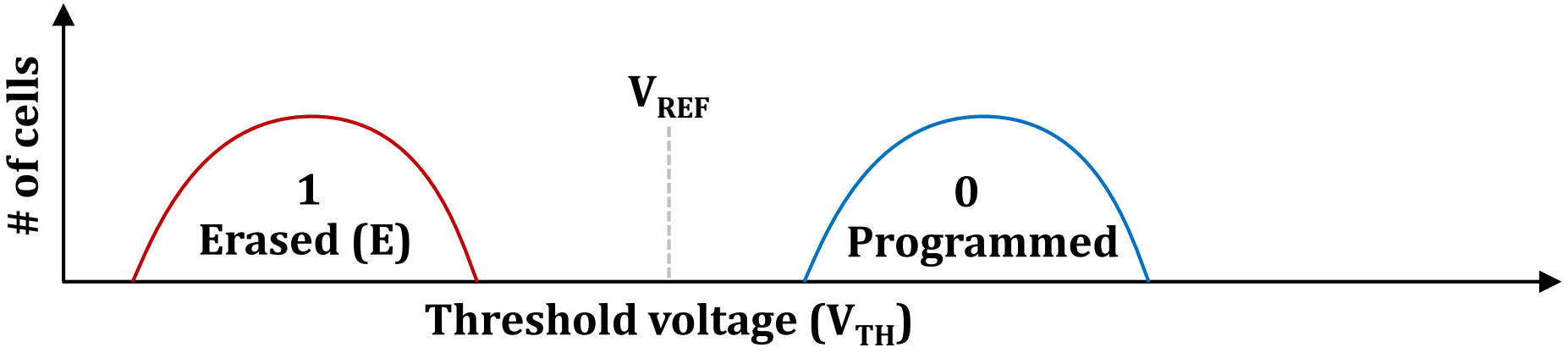




Basic Operation: Page Read

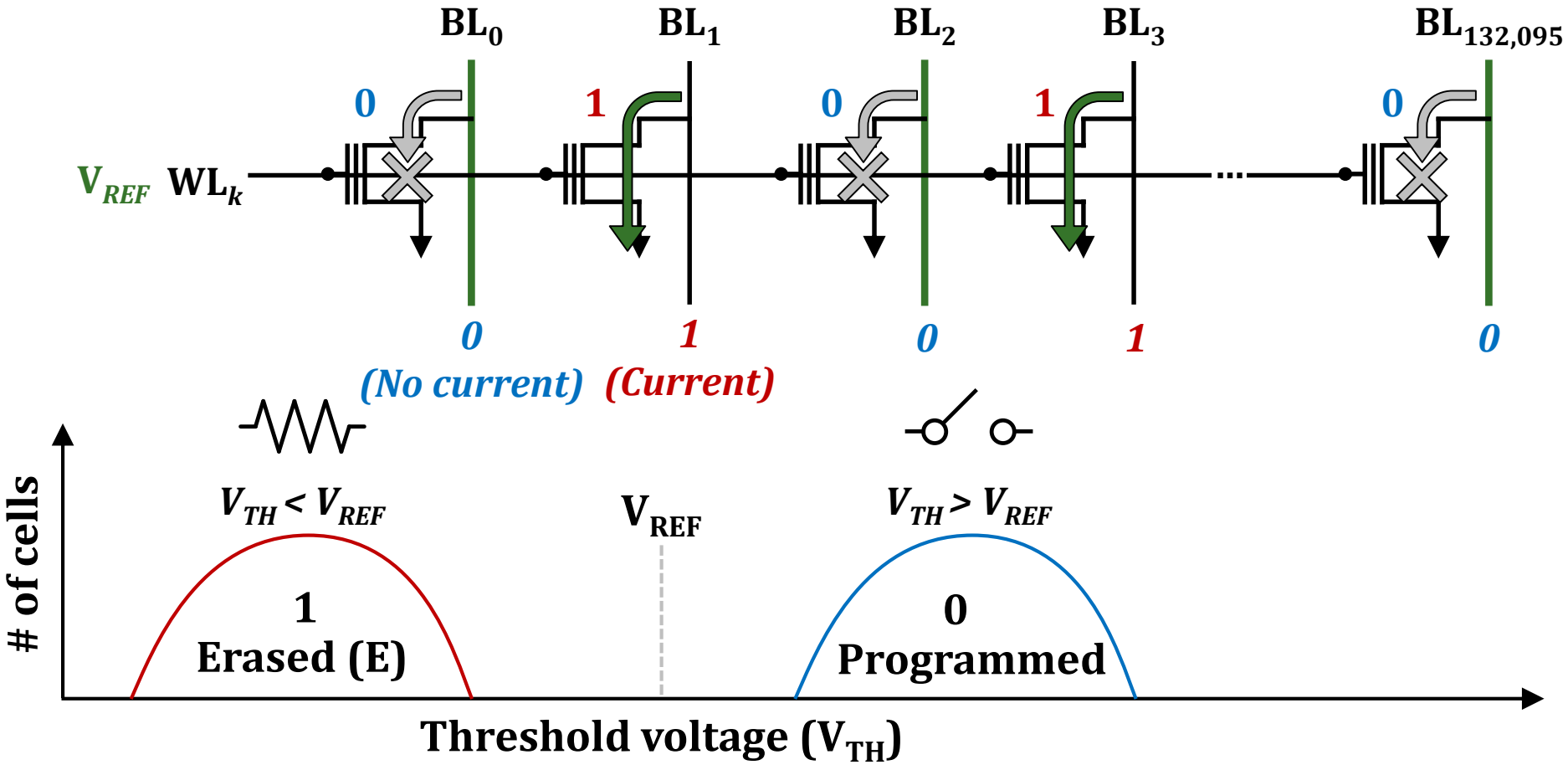
- BL control – Charge all BLs





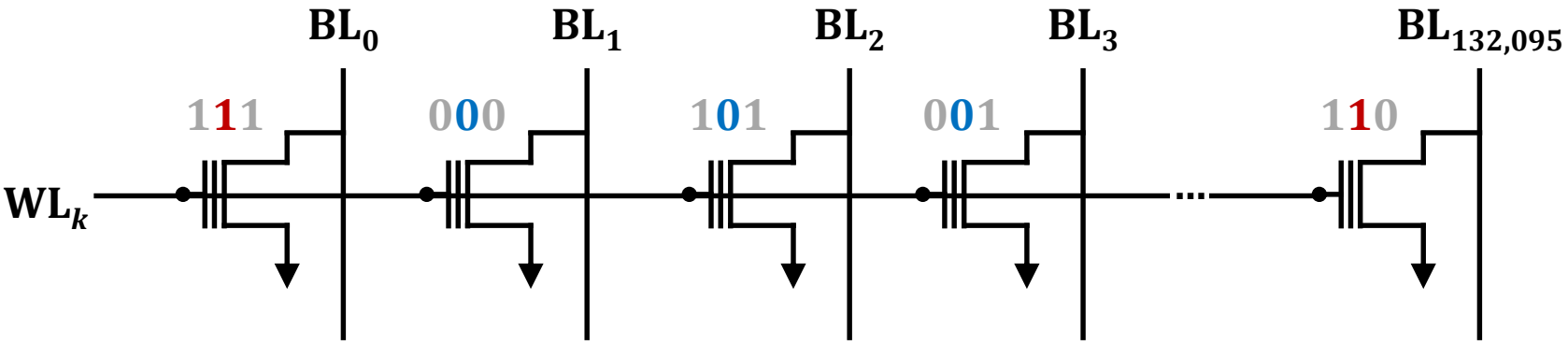
Basic Operation: Page Read

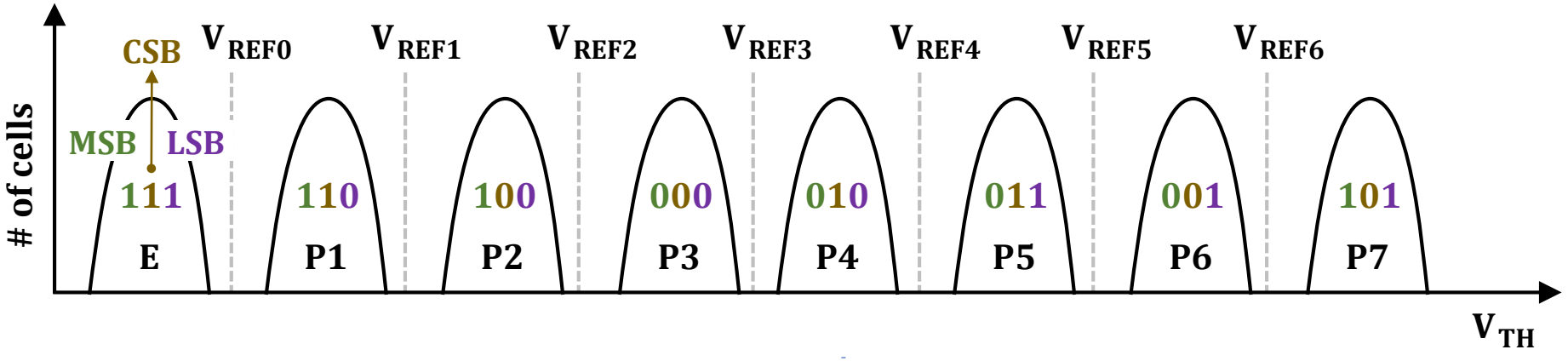
- Sensing the current through BLs



Basic Operation: Page Read - MLC

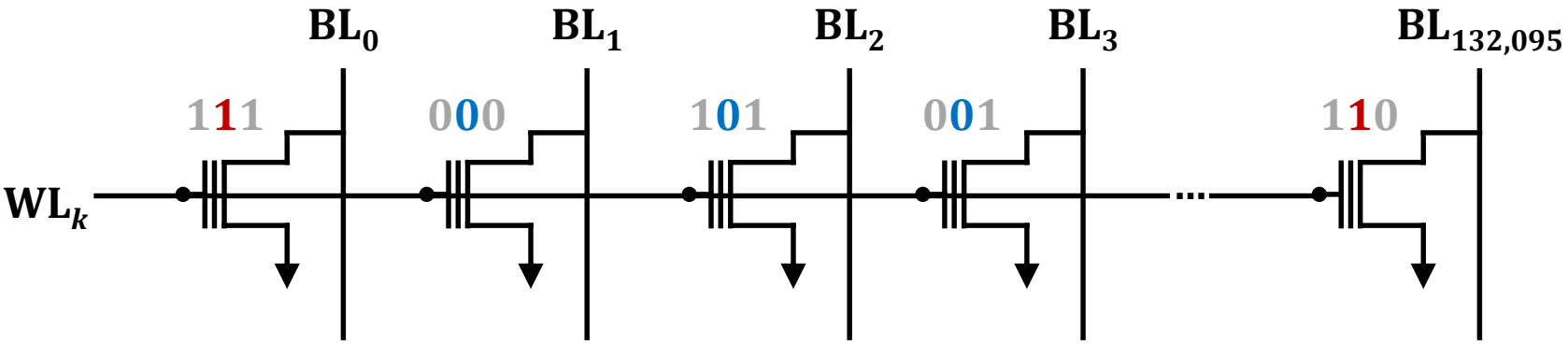
- Sensing the current through BLs

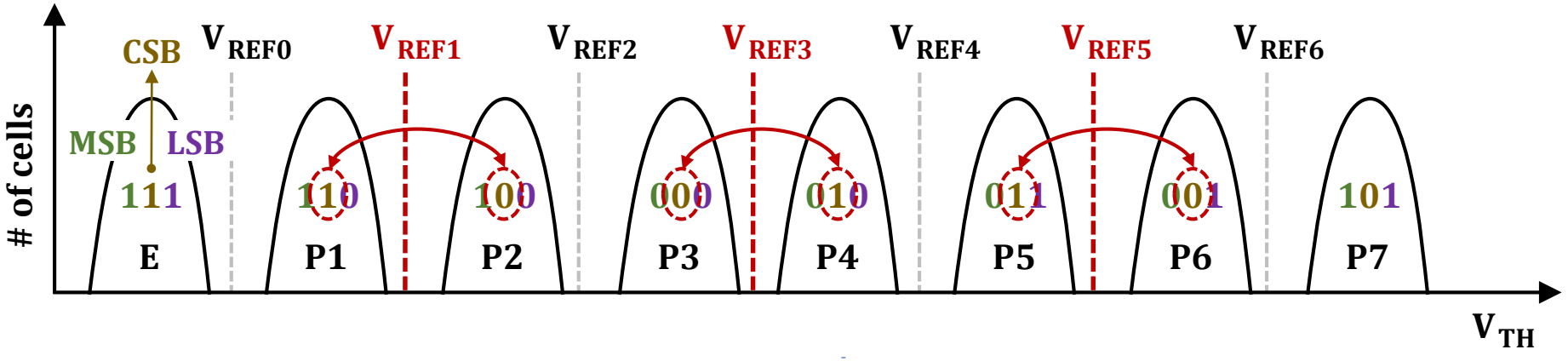




Basic Operation: Page Read - MLC

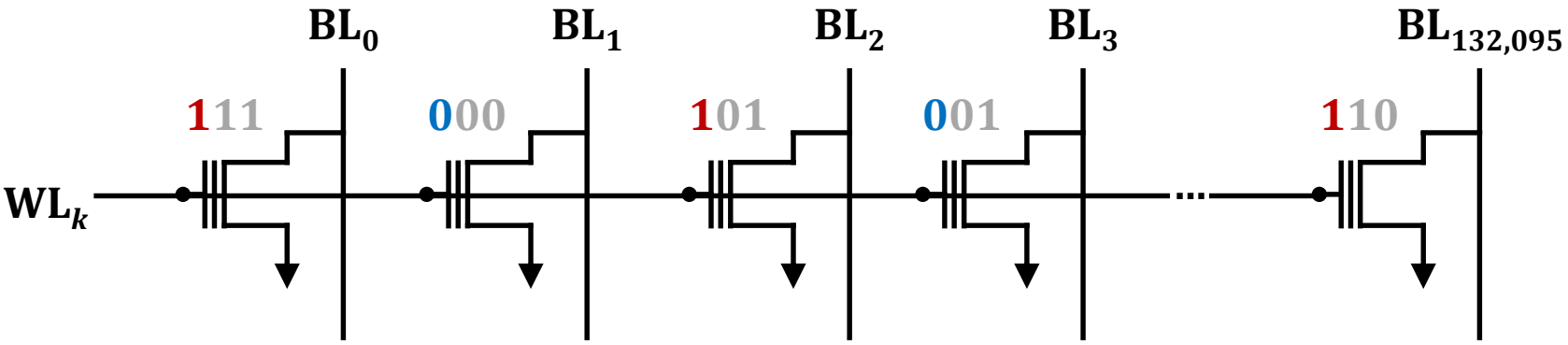
- Sensing the current through BLs

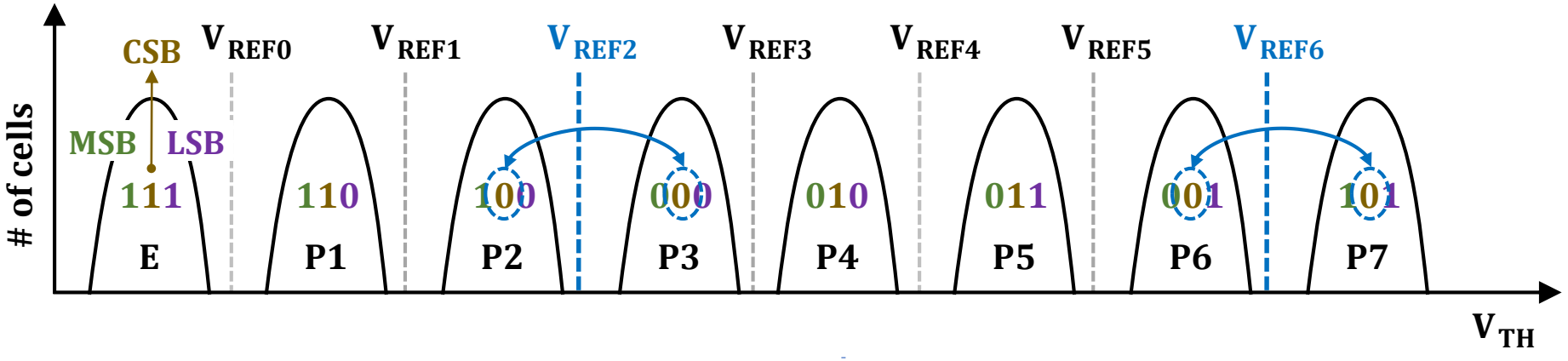




Basic Operation: Page Read - MLC

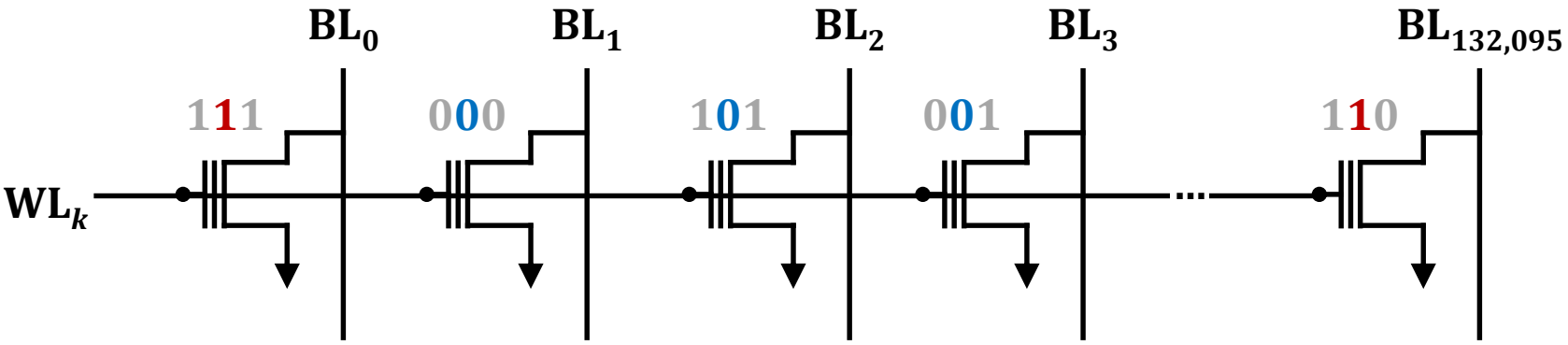
- Sensing the current through BLs

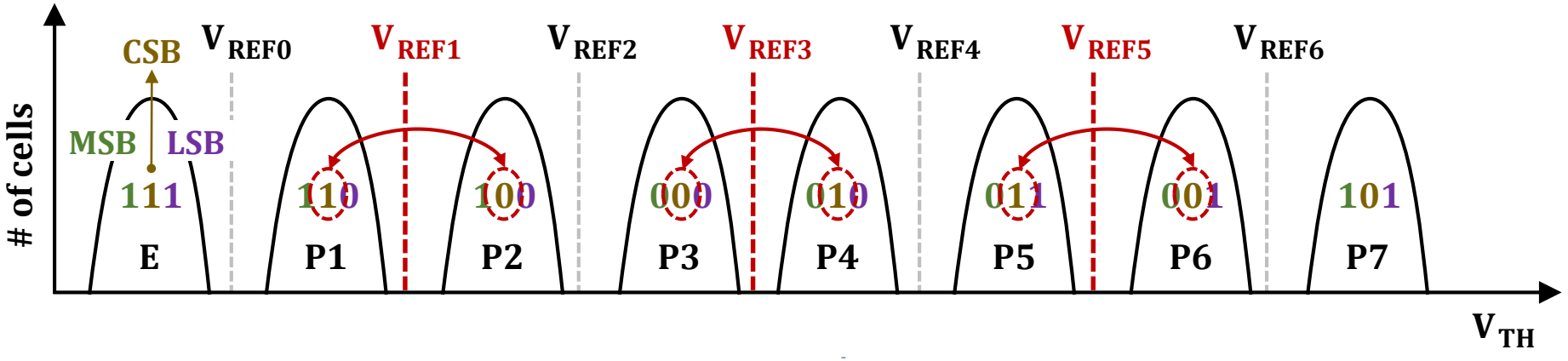




Basic Operation: Page Read - MLC

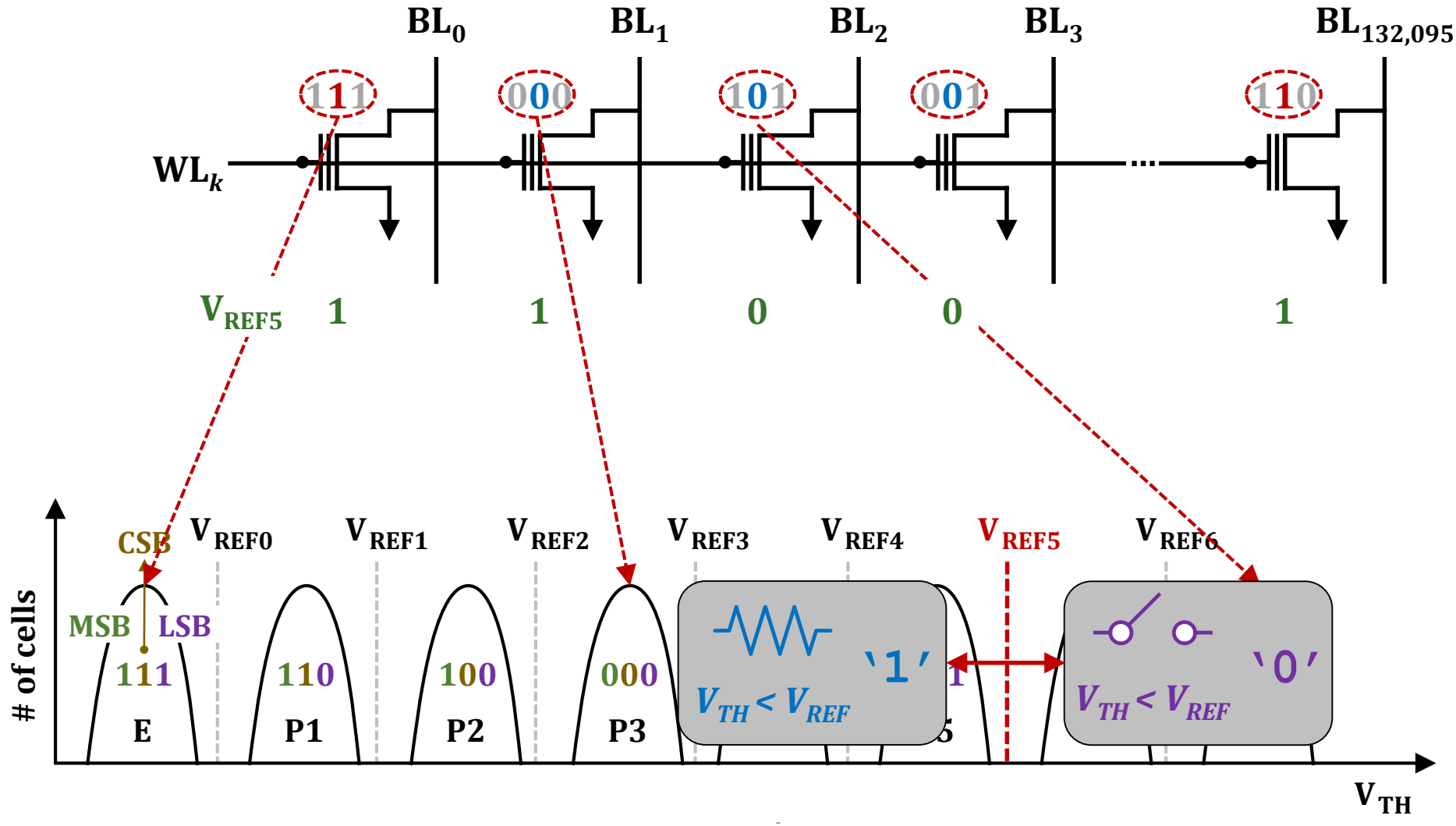
- Sensing the current through BLs





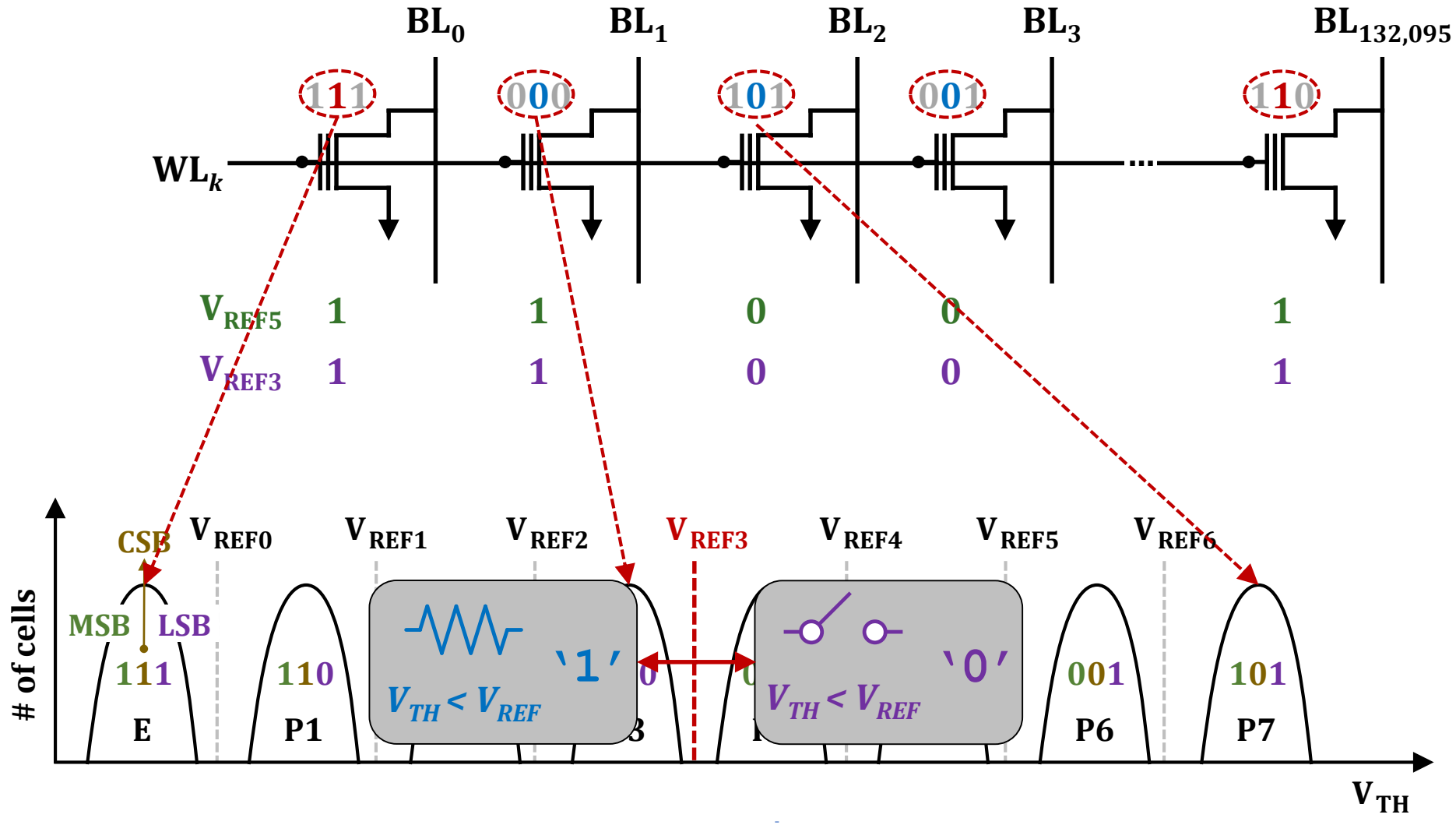
Basic Operation: Page Read - MLC

- Sensing the current through BLs



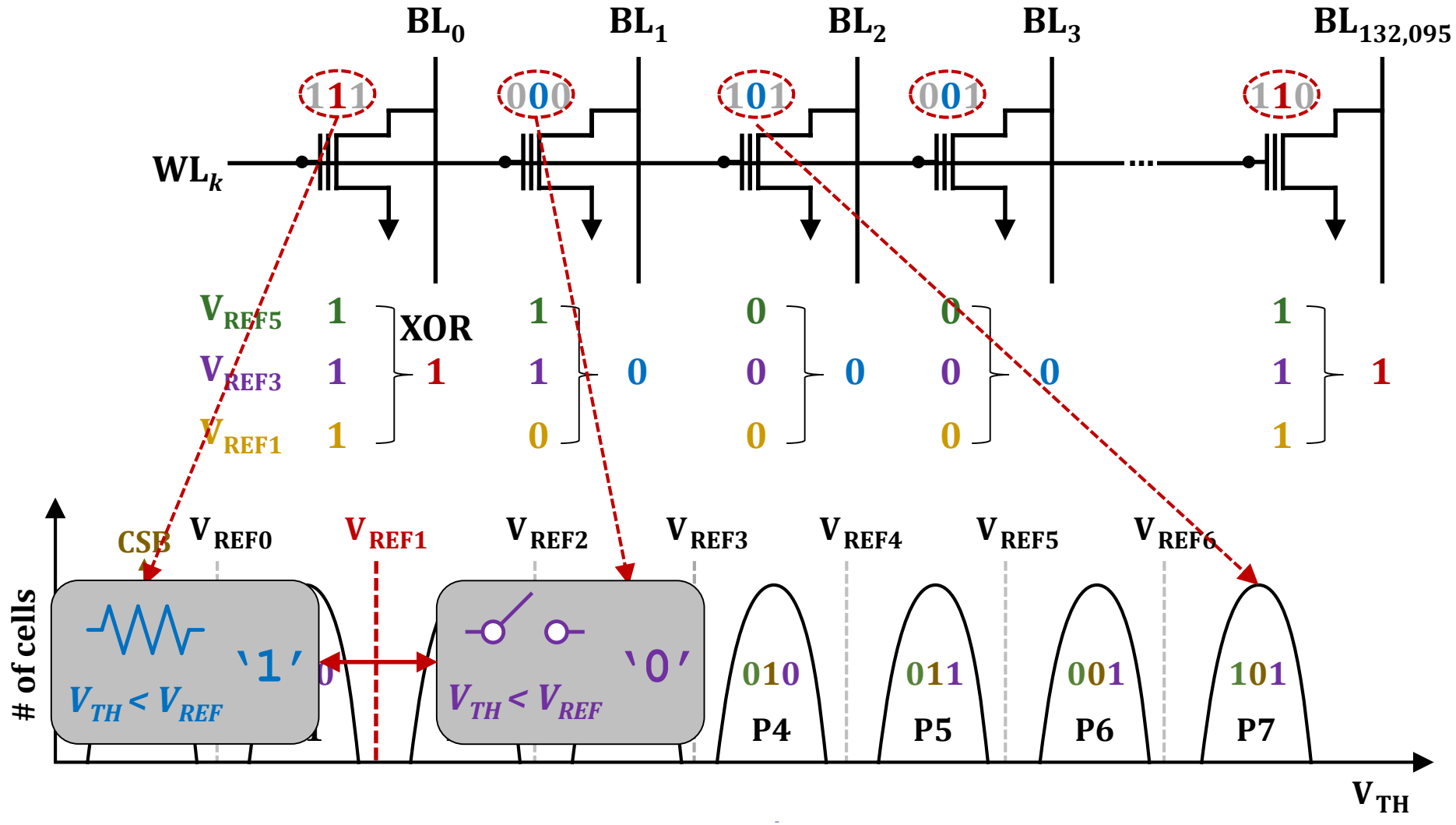
Basic Operation: Page Read - MLC

- Sensing the current through BLs



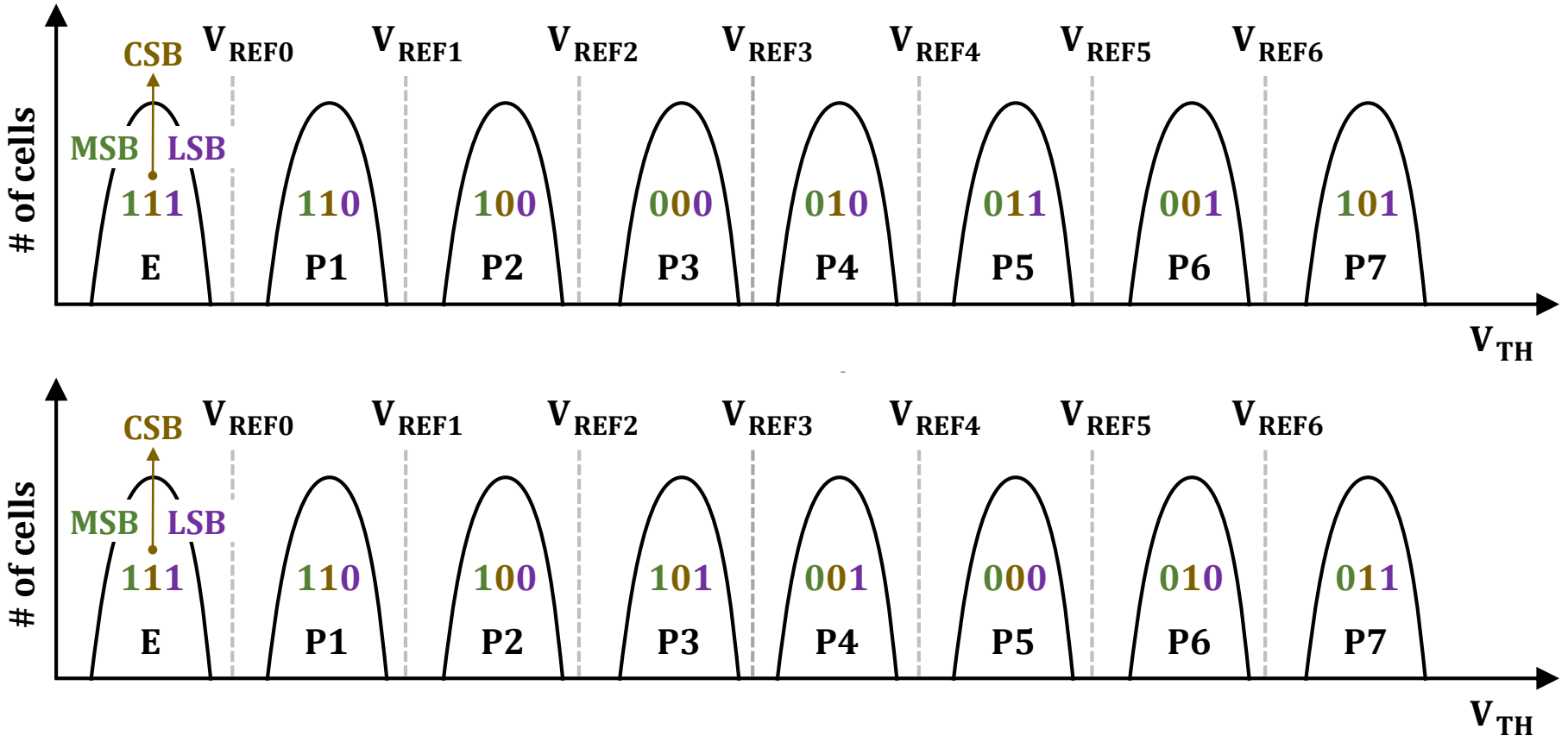
Basic Operation: Page Read - MLC

- Sensing the current through BLs



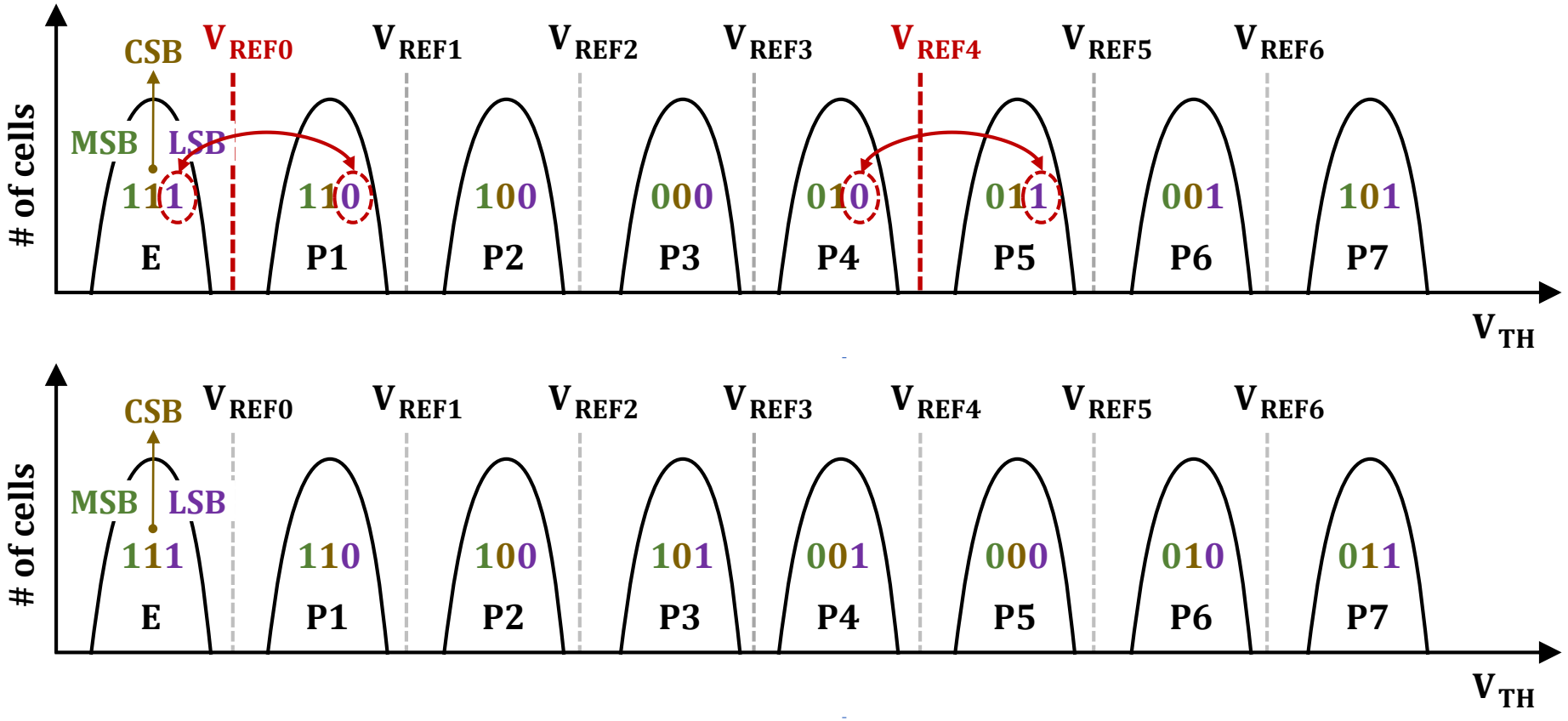
Basic Operation: Page Read – Takeaways

- Bit-encoding affects the read latency!
  - Compare # of sensing for LSB



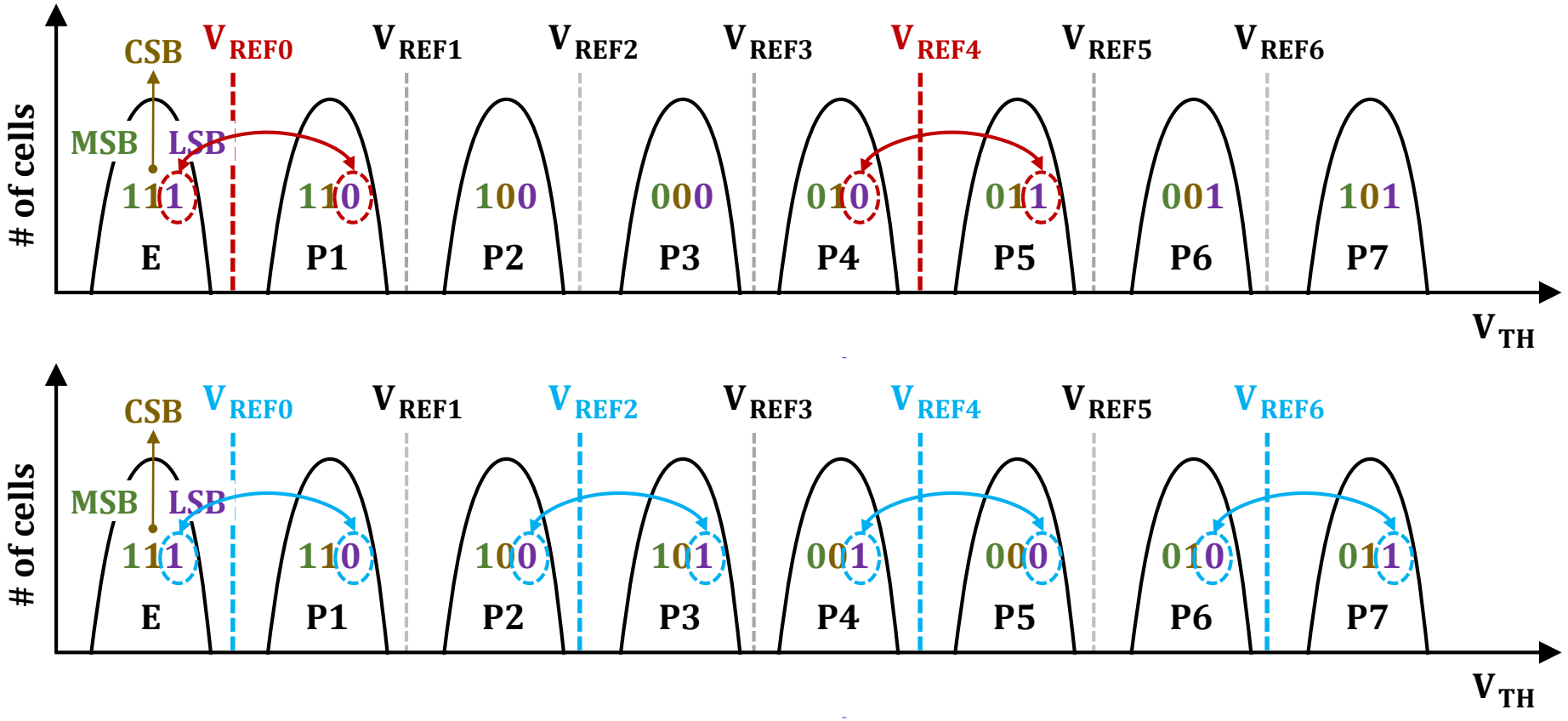
Basic Operation: Page Read – Takeaways

- Bit-encoding affects the read latency!
  - Compare # of sensing for LSB



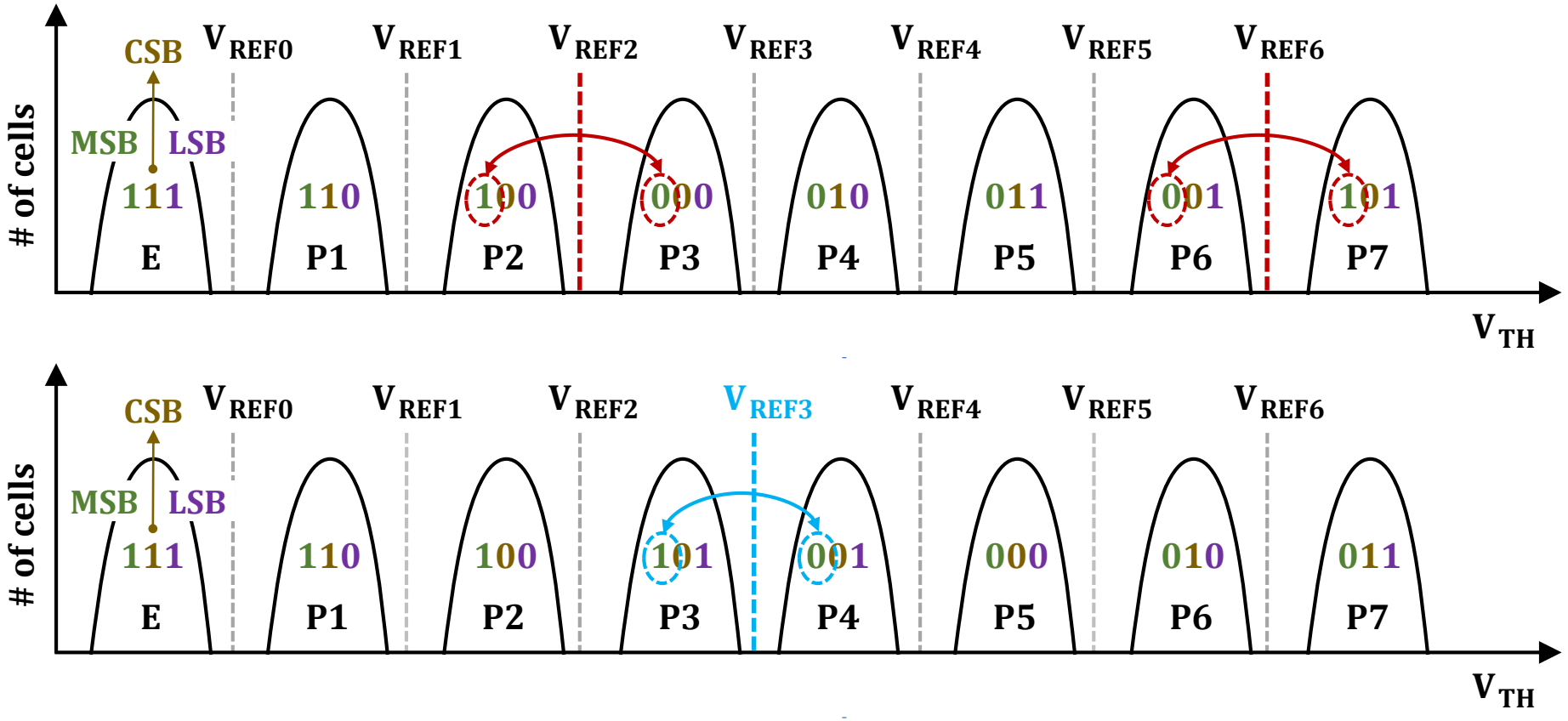
Basic Operation: Page Read – Takeaways

- Bit-encoding affects the read latency!
  - Compare # of sensing for LSB



Basic Operation: Page Read – Takeaways

- Bit-encoding affects the read latency!
  - Compare # of sensing for LSB

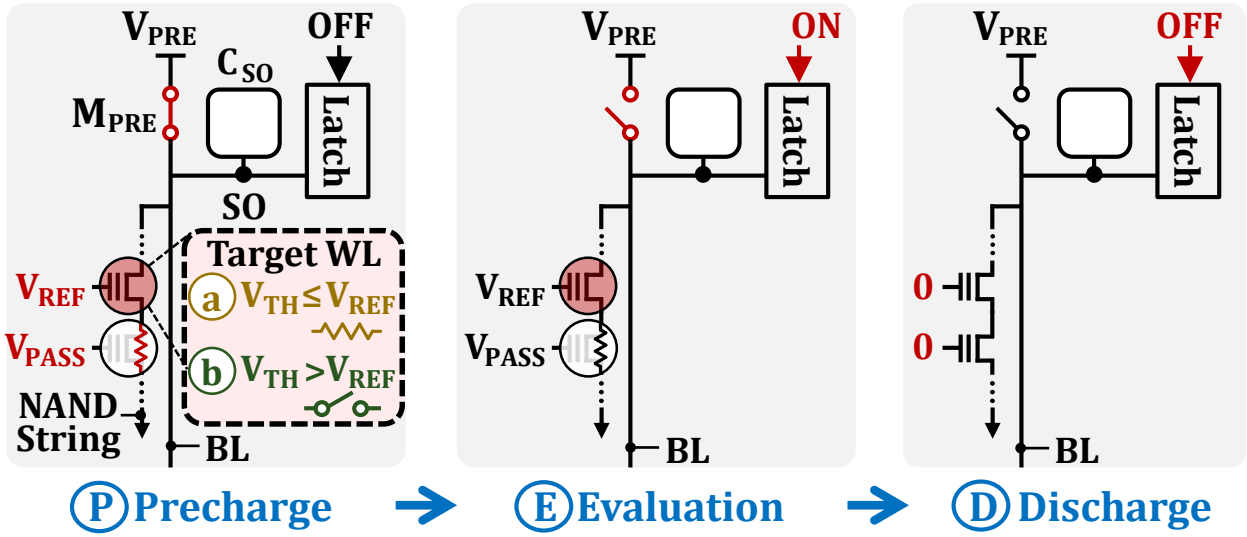


Today's Agenda

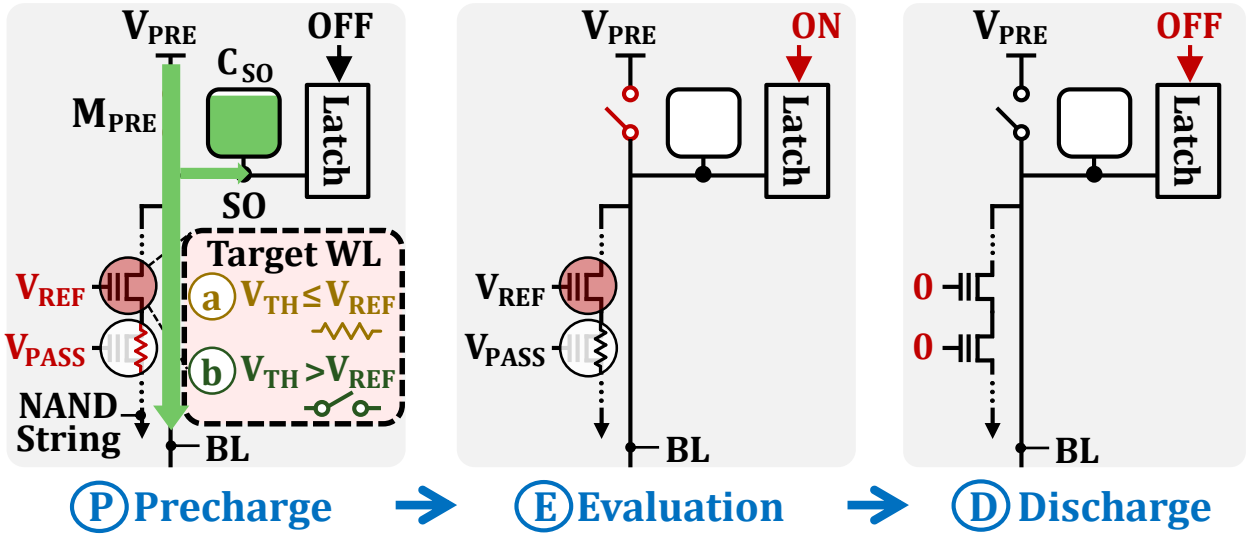
- NAND Flash Page Read/Write
- Sensing Circuitry

Read Mechanism

- NAND flash read mechanism consists of three steps:
  - 1) Precharge
  - 2) Evaluation
  - 3) Discharge

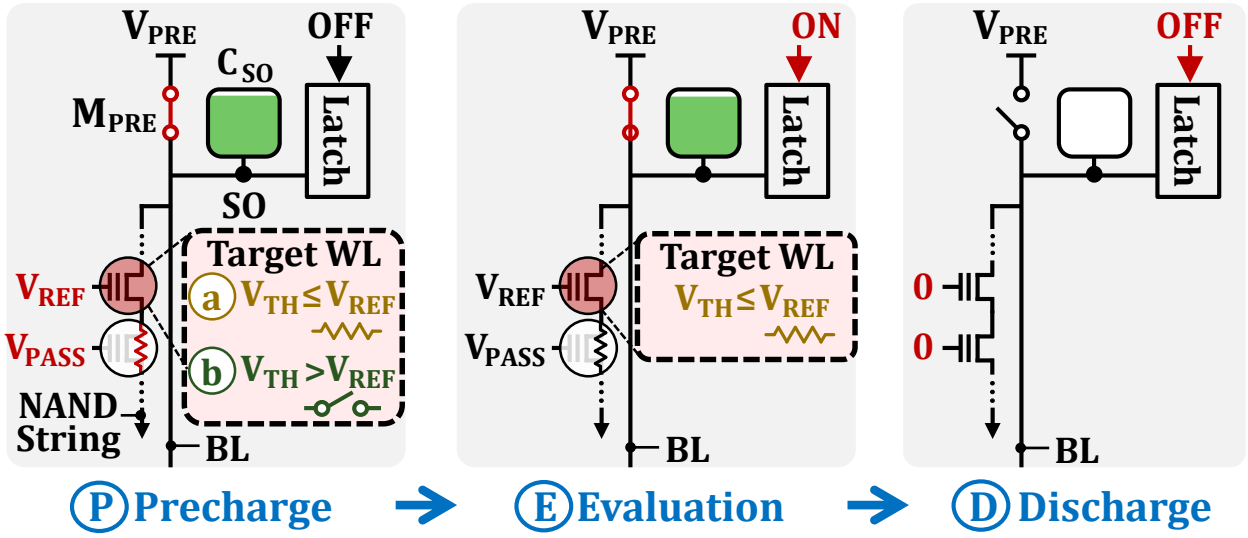


Read Mechanism: Precharge

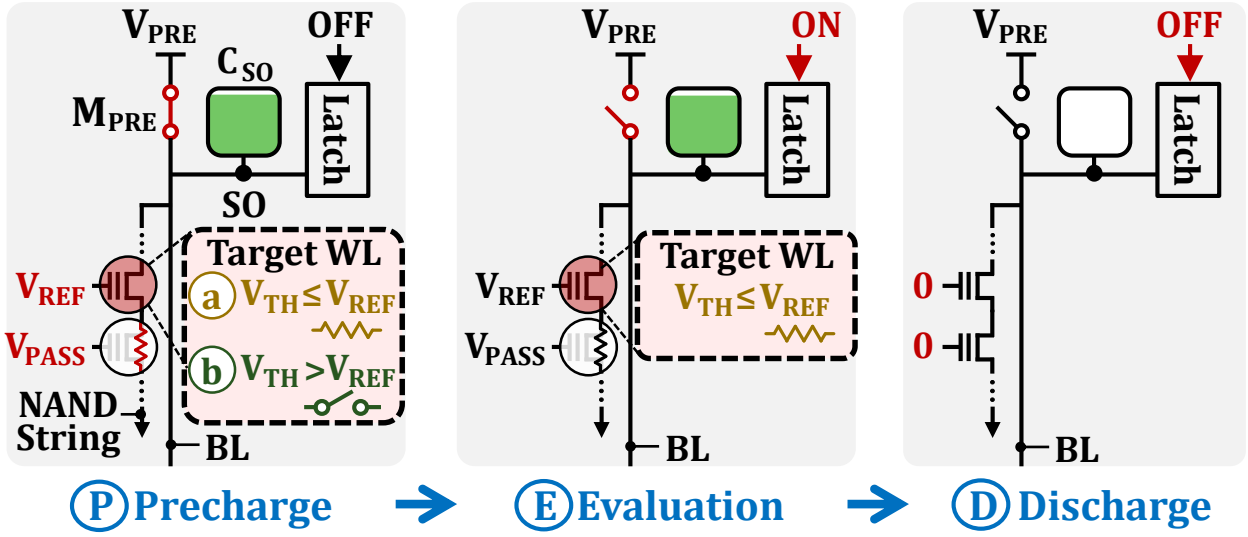


Enable precharge transistor M_{PRE} to charge all target BLs and their sense-out capacitors (C_{SO}) to V_{PRE}

Read Mechanism: Evaluation

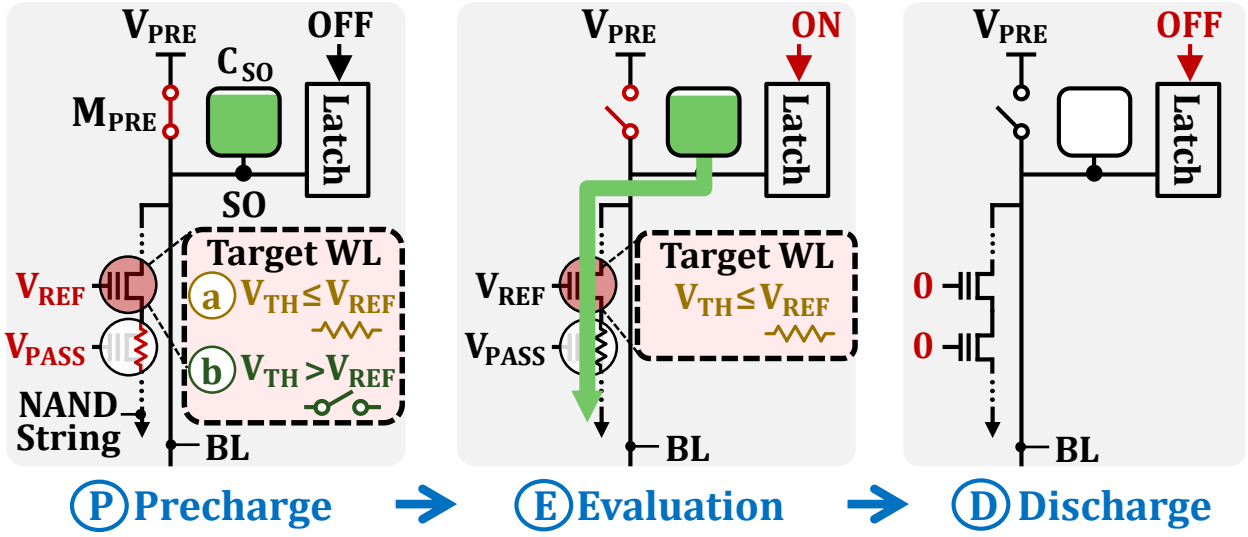


Read Mechanism: Evaluation



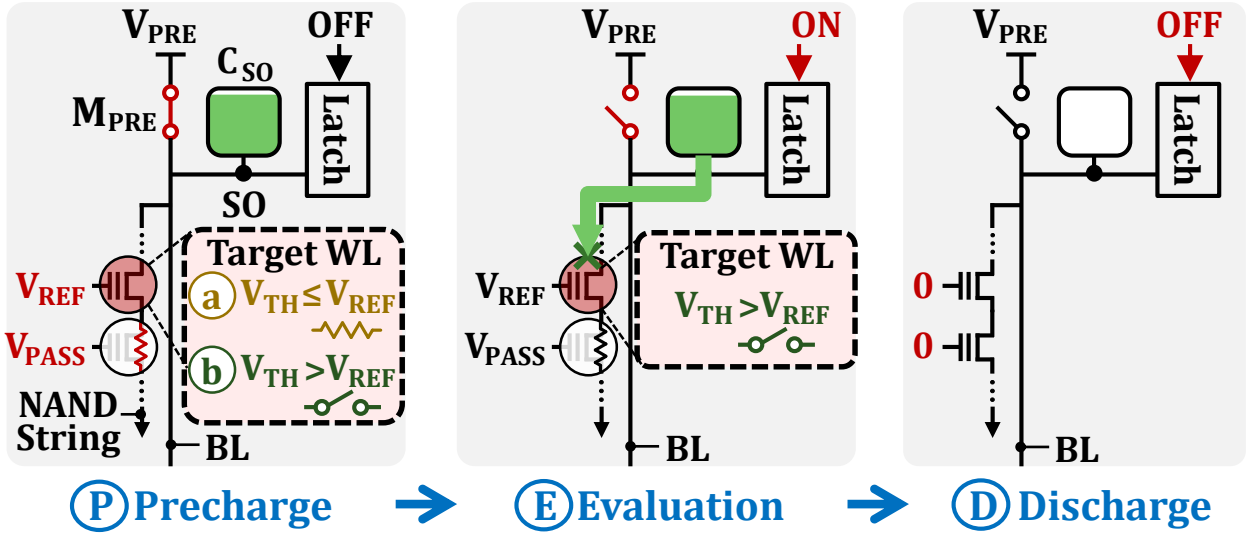
Disconnect the BLs from V_{PRE} and enable the latching circuit

Read Mechanism: Evaluation



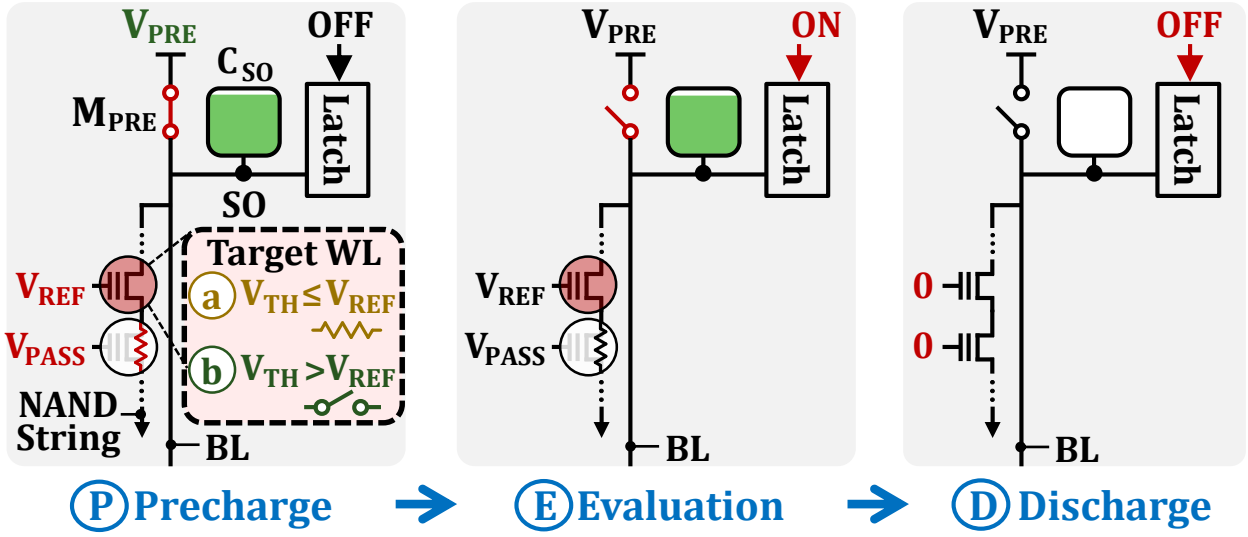
If $V_{TH} \leq V_{REF}$, the charge in C_{SO} quickly flows through the NAND string (Sensed as 1)

Read Mechanism: Evaluation



If $V_{TH} > V_{REF}$, the target cell blocks the BL discharge current (Sensed as 0)

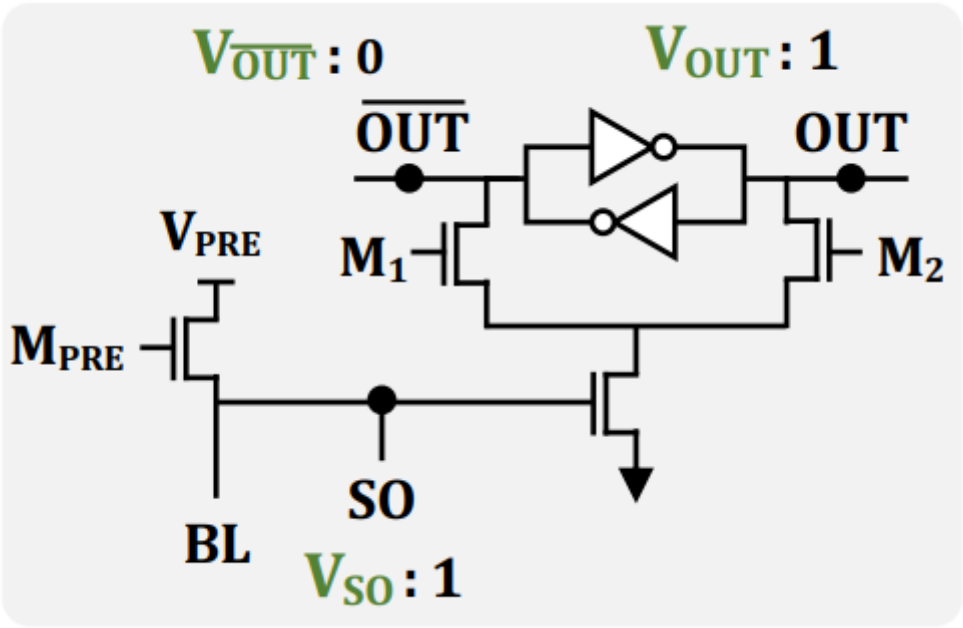
Read Mechanism: Discharge



Bitlines are discharged to return the NAND string to its initial state for future operations

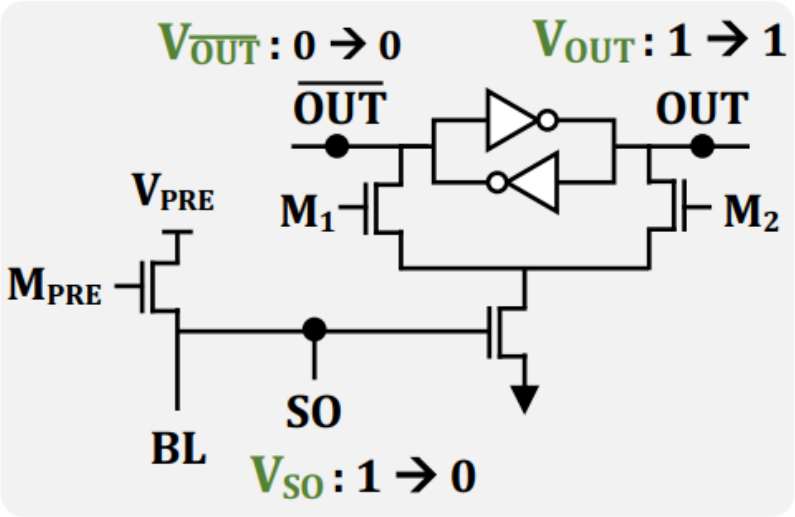
Latching Circuit

- Before the evaluation step, the chip initializes the latching circuit
  - Activating transistor M_1
  - $V_{\overline{OUT}} = 0$
  - $V_{out} = 1$

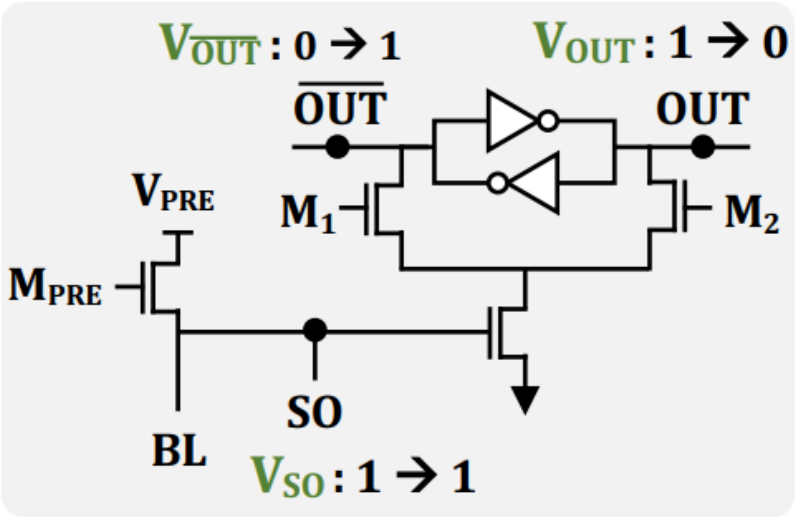


Latching Circuit

- The evaluation step
  - Disables M_{PRE} and M_1
  - Enables M_2



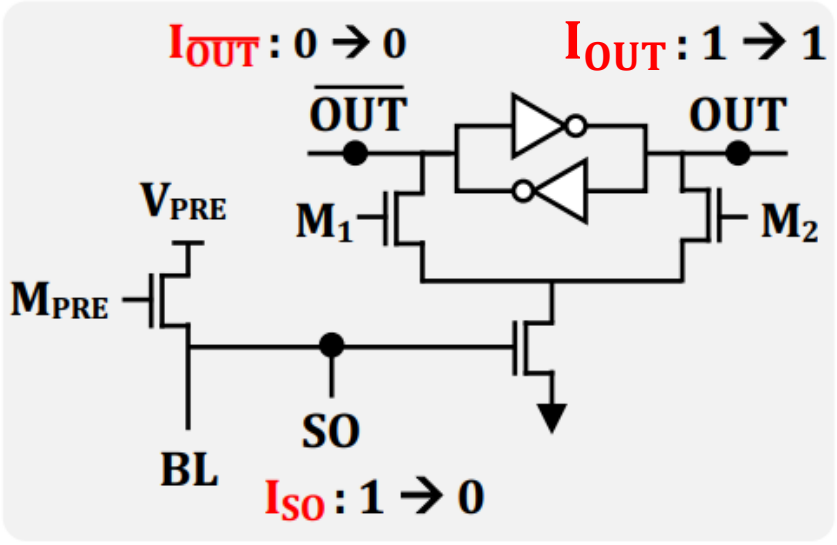
(a) $V_{TH} \leq V_{REF}$



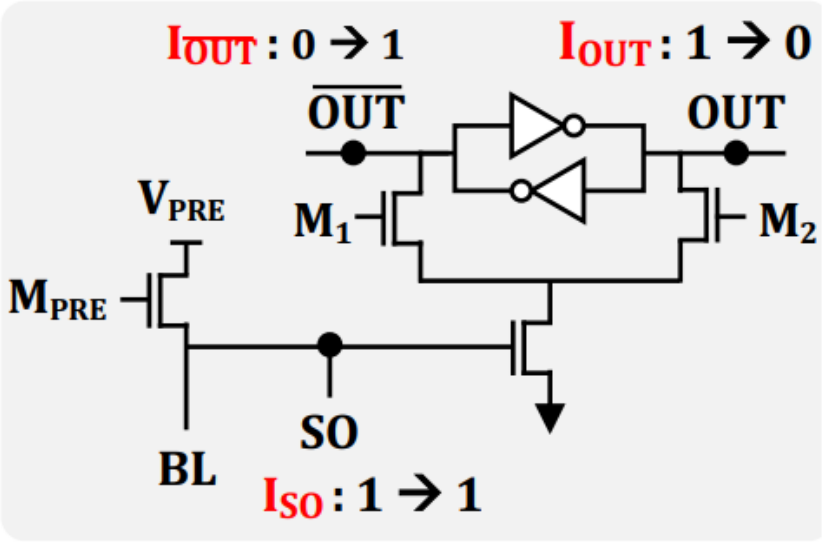
(b) $V_{TH} > V_{REF}$

Inverse Read

- Performing an inverse read by simply changing the activation sequence of M_1 and M_2
  - The precharge step activates M_2
  - The evaluation step disables M_2 and activates M_1



(a) $V_{TH} \leq V_{REF}$



(b) $V_{TH} > V_{REF}$

Flash-Cosmos, MICRO'22

Flash-Cosmos: In-Flash Bulk Bitwise Operations Using Inherent Computation Capability of NAND Flash Memory

Jisung Park[§▽] Roknoddin Azizi[§] Geraldo F. Oliveira[§] Mohammad Sadrosadati[§]
Rakesh Nadig[§] David Novo[†] Juan Gómez-Luna[§] Myungsuk Kim[‡] Onur Mutlu[§]

[§]*ETH Zürich* [▽]*POSTECH* [†]*LIRMM, Univ. Montpellier, CNRS* [‡]*Kyungpook National University*

https://arxiv.org/abs/2209.05566.pdf

Required Material

- Yu Cai, Saugata Ghose, Erich F. Haratsch, Yixin Luo, and Onur Mutlu,
  "Errors in Flash-Memory-Based Solid-State Drives: Analysis, Mitigation, and Recovery,"
  Invited Book Chapter in Inside Solid State Drives, 2018
  - Introduction and Section 1

- Jisung Park, Myungsuk Kim, Myoungjun Chun, Lois Orosa, Jihong Kim, and Onur Mutlu,
  "Reducing Solid-State Drive Read Latency by Optimizing Read-Retry," In ASPLOS, 2021

Recommended Material

- Arash Tavakkol, Mohammad Sadrosadati, Saugata Ghose, Jeremie Kim, Yixin Luo, Yaohua Wang, Nika Mansouri Ghiasi, Lois Orosa, Juan Gómez Luna, and Onur Mutlu, "FLIN: Enabling Fairness and Enhancing Performance in Modern NVMe Solid State Drives," In ISCA, 2018

P&S Modern SSDs

NAND Flash Read/Write Operations

Dr. Mohammad Sadrosadati
Prof. Onur Mutlu
ETH Zürich

Fall 2022

19 October 2022